Movement-based Human-Machine Collaboration: a Human-centred AI approach (accreditation to supervise research)

Sotiris Manitsaris

Sorbonne Université

MÉMOIRE D'

HABILITATION

À DIRIGER DES RECHERCHES

Spécialité : Sciences de l'Ingénieur

*THESIS FOR **ACCREDITATION TO SUPERVISE RESEARCH***

Specialisation: Engineering Sciences

Movement-based Human-Machine Collaboration: *a Human-centred AI approach*

defended by

Sotiris **Manitsaris**

prepared at

the Centre for Robotics, Mines Paris, PSL Université Paris

defended on the 6th May 2021

Reviewers:	
Emilios **Cambouropoulos**, Professor	- Aristotle University of Thessaloniki
Sarah **Fletcher**, Senior Researcher	- University of Cranfield
Wendy **Mackay**, Research Director	- INRIA
Evaluators:	
Brigitte **d'Andréa-Novel**, Professor	- IRCAM, Sorbonne Université
Amel **Bouzeghoub**, Professor	- Télécom Paris Sud
Leontios **Hadjileontiadis**, Professor	- Khalifa University
Patrick **Henaff**, Professor	- Mines Nancy, Université de Lorraine
Fabien **Moutarde**, Professor	- MINES ParisTech, PSL Université Paris

Movement-based Human-Machine Collaboration: a Human-centred AI approach

Abstract: The context of this thesis is the collaboration between humans and machines in various industrial real-world situations. I propose collaboration mechanisms that are based on Human-centred Artificial Intelligence, which I define as methods and concepts of machine learning and pattern recognition on signals recorded from the human body. I am interested in enabling human-machine partnerships in which the machine can understand and anticipate the human gestures and actions and react accordingly. Two scientific and technological hypotheses have oriented my research in movement-based Human-Machine Collaboration: 1. whether a machine can learn to recognize kinematic parameters of situated expert and non-expert gestures; and 2. whether gesture recognition can be used as an alternative to instrumental interaction mechanisms. These hypotheses were confirmed through a number of tests and experiments conducted on Human-Robot Collaboration, computer-mediated sensori-motor human learning and Digital Musical Instruments.

Keywords: machine, collaboration, action, gesture, movement, machine learning, interaction, partner, stochastics, robotics, sensori-motor learning, Digital Musical Instrument (DMI), Artificial Intelligence (AI)

Contents

Chapter 1

Introduction

Contents

1.1 Human-centred Artificial Intelligence

AI was founded as an academic discipline in 1958, but it is only in recent years that we have been able to use and take advantage of it, in its various forms, in business and industrial sectors and in our everyday lives. Even so, it has to be said that the advances in human sensing and AI over the past 20 years have not always been matched by equivalent advances in natural interaction and collaboration with machines. Fractured interfaces have been developed around closed systems (e.g. robotics, vehicles, Internet of Things (IoT)s etc.). In most cases, they are organised through physical instrumental interactions that make use of intermediary mechanisms (i.e. buttons etc.).

By introducing gestures as a modality for interaction, the digital world has taken a step forward towards less 'rigid' machines. Nevertheless, machines still lack the necessary layers of perception that would permit them to understand situated human movements and adapt their behaviour accordingly and thus enable them to collaborate as real partners. For example, a lot of progress has been made in collaborative robotics, especially in the training phase of the robot (e.g. learning by demonstration) but smooth Human-Robot Collaboration, which would allow for complementarity of skills, nevertheless remains a challenge.

I define as Human-centred Artificial Intelligence (HAI) concepts and methods of machine learning and pattern recognition on signals recorded from the human body. In this thesis, I focus, in particular, on HAI paradigms, that are relevant to situations where humans collaborate with intelligent machines. Throughout my various research projects, I have always been interested in how intelligent machines can understand human body movements, collaborate with them and adapt their behaviour accordingly. Having worked on a wide range of HAI research projects, I have had the opportunity to study human movement in a variety of real-life situations: professional gestures and body actions in manufacturing, workers' collaboration with robots, traditional craftsmanship for human learning of movement skills, musical gestures for developing digital musical instruments, as well as contactless hand and finger interactions with intelligent and automated vehicles.

1.2 Scientific and technological hypotheses

Situated cognition is a theory that posits that knowing is inseparable from doing by arguing that all knowledge is situated in activity bound to social, cultural and physical contexts [Anderson *et al.* 1996]. The elaboration, structuring and performance of gestures is naturally connected with their specific context and thus their extraction from the situated environment might seriously impact the gestural phenomenon itself.

In my research, I use a human-centred interactive model to represent situated body actions, where the human is both a trigger and transmitter connecting the perception (*mind/environment interaction*), the knowledge (*understanding of a process*) and the gesture (*movement skills*), as shown in Figure 1.1. For example, in the professional context of pottery-throwing, the human has the theoretical knowledge of the properties of clay, she perceives the fact that her clay is revolving, together with the round plate, and she successfully performs micro-adjustments over time in order to centre the clay on the plate, by applying the appropriate finger motions.

Machines are '*systems of intelligence*' that, in most cases, use human sensing technologies, whether embedded in them or external to them, in order to perceive the human presence and/or motion. In the human-centred interactive model, machines can constitute the *goal*, i.e. movement-based interaction between the driver and the dashboard, or the *means* for interaction, i.e. computer-mediated human learning of movement skills.

By way of leveraging out the above conclusions, I have oriented my research towards the development of a methodology that allows for movement-based Human-Machine Collaboration. Consequently, a number of scientific questions have emerged naturally and have driven my research:

Q1: *Can machines learn to recognize kinematic parameters of situated expert and non-expert gestures?*

In order to confirm or refute this scientific question, a number of tests were conducted through various experiments, derived from different scenarios, to check whether machine learning and dynamic pattern recognition are able to recognize the stochasticity and variability of gestures over time.

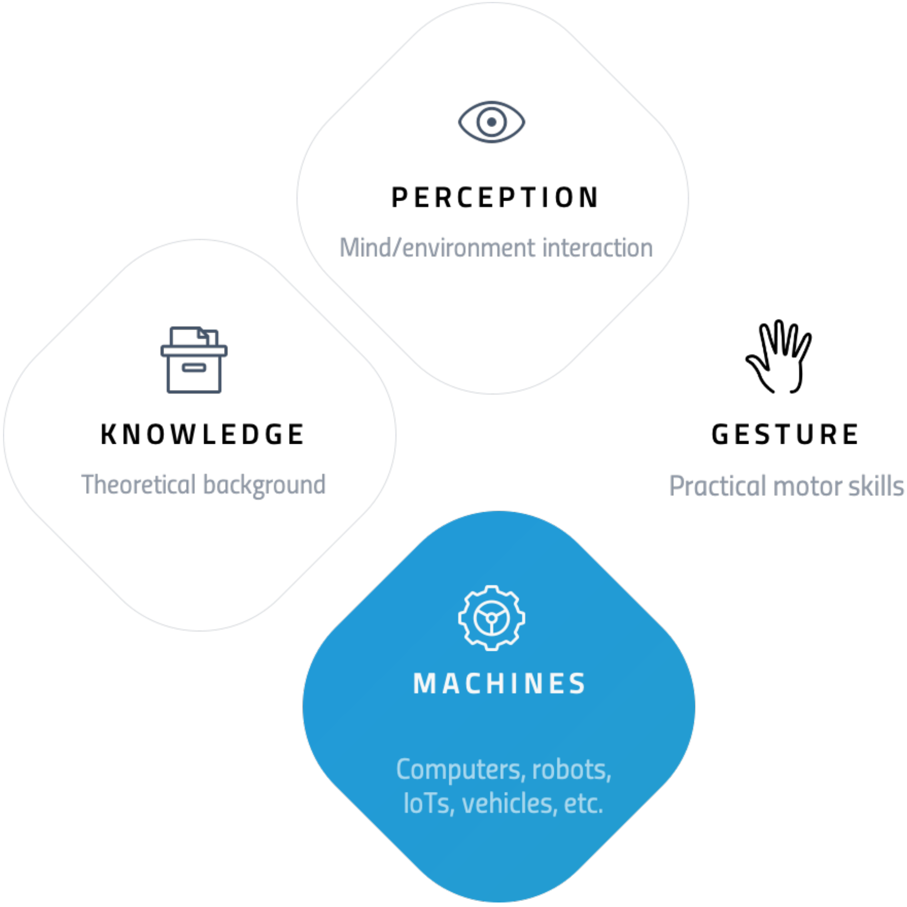


Figure 1.1: A human-centered model for movement-based collaboration with systems of interaction

Q2: *Can gesture recognition be used as an alternative to instrumental interaction mechanisms?*

To answer this scientific question, a number of tests were conducted in order to check whether gestures, once recognized by machines, can be used as an alternative modality to intermediary instrumental mechanisms that can achieve a more natural collaboration.

The above scientific and technological questions are split into other subquestions that depend on the specificities of each type of gesture as well as the situation of their execution and the use-case in general.

1.3 Industry-oriented research

Following the strategic model of the Centre for Robotics and MINES ParisTech (Ecole des Mines de Paris), an Engineering School of PSL Université Paris, I had the opportunity to perform experiments and develop technological prototypes that were derived from industry-oriented research. The close links of MINES ParisTech with industry, but also the opportunity to participate in European collaborative research and innovation projects, generated favourable conditions for studying Human-Machine Collaboration extensively.

From a social point of view, 'gestures are everywhere' in our life. This is also valid in our professional environments, whether industrial or not. Living in an 'Ubiquitech' era where, on the one hand, technology is becoming embedded everywhere and in everything, yet on the other hand, the Fourth Industrial Revolution (Industry 4.0) 'has yet to grab hold on a grand scale' (the Stall Zone within an *S-Curve*), industry faces big challenges concerning how humans can cooperate efficiently with the 'new intelligent machines' [Frank *et al.* 2017].

For this reason, the main methodological framework, the technological prototypes, together with the scientific questions that I have worked on, have been motivated and driven by the above conclusions. This means that my major contributions address mainly scientific and technological challenges in the three sectors of Factory of the Future (FoF), Creative and Cultural Industries (CCI) and Vocational Training.

1.3.1 Factory of the Future

The main scientific axes of my contributions in this field are summarized in:

- *professional gesture recognition for Human-Robot Collaboration in automotive assembly lines.* The results have been published mainly in the scientific communities of Autonomous Robots (Springer), Ro-Man (IEEE), AI and Robotics (Frontiers) and Movement and Computing (MOCO).

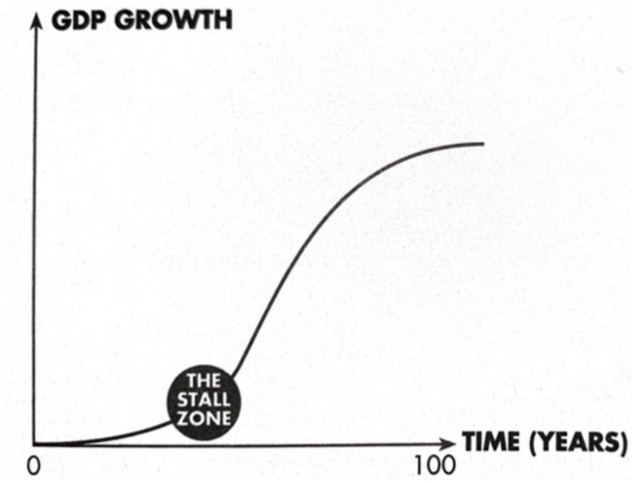


Figure 1.2: The *S-Curve* and the Stall Zone [Frank *et al.* 2017]

- *monitoring of the ergonomic performance of the operators* (ongoing work). The results have been published mainly in the scientific communities of Human Factors in Applied Ergonomics (Elsevier) and Ambient Intelligence and Humanized Computing (Springer).

The research I have conducted has been funded for the main part by the Industrial Chair 'PSA Peugeot Citroën: Robotics and Virtual Reality' as well as the ongoing Horizon 2020 'Collaborate' project.

1.3.2 Creative and Cultural Industries

The main scientific axis of my contributions in this field can be found, in summary, in:

- *musical gesture recognition for developing a Digital Musical Instrument.* The results have been published for the most part in the NIME (New Interfaces for Musical Expression) and MOCO scientific communities and have contributed to the creation of the 'Embodme' Company, SpinOff MINES ParisTech.

The research I conducted was financially supported by the FP7 'i-Treasures' project and the Greek national 'ArtiMuse' project of 'Research Excellence'.

1.3.3 Vocational Training

The main scientific axis of my contributions in this field can be found in:

- *movement skills transmission in manual professions and the Intangible Cultural Heritage.* The results are published in the scientific communities of Computing and Cultural Heritage (ACM), Technology Enhanced Learning and Intelligent Systems (IEEE).

The research I conducted was funded mainly by the FP7 'i-Treasures' project, the Greek national 'ArtiMuse' project of 'Research Excellence' and the Carnot M.I.N.E.S. project 'SMART'.

1.3.4 Other industrial sectors

In addition to the above domains, I have also made peripheral contributions to the industrial sectors of Intelligent Vehicles as well as to Defence and Security that are summarized below and are very briefly presented in Chapter 6:

- *vision-based contactless driver-vehicle interaction for infotainment.* The results have been published for the most part within the Computer Vision community.

The research I conducted was funded mainly through a direct contract with the PSA Group, as a PhD supervision.

- *hand and finger gestural control of a Unmanned Aerial Vehicle (UAV).* The results have been published mainly in Human Factors in Applied Ergonomics (Elsevier) and the Human Factors in Computing Systems (ACM CHI).

The research I conducted was funded mainly through a direct contract with the SAFRAN Group, as a PhD supervision.

1.4 Structure of the thesis

The goal of Chapter 2 is to present a generic methodology for gestures and actions recognition that is based on a stochastic-biomechanic modelling of the spatio-temporal dynamics of the human body. The proposed algorithm is validated through various datasets of human performances. It demonstrates the feasibility of recognizing situated professional gestures from various industries. In the following chapters, I demonstrate the feasibility of converting a machine to a partner through movement-based interactions and HAI-driven methodologies for a robot (Chapter

3), a computer (Chapter 4) and a DMI (Chapter 5), at the same time pursuing the research questions that I presented in Section 1.2. In Chapter 3, I present a number of experiments in recognizing professional gestures of operators when collaborating with robots in the assembly lines of the automotive industry (FoF), in which the robot is approached as a partner of the operator instead of a tool. In Chapter 4, a methodology for training the computer to assist the human in sensori-motor learning of movement skills in craftsmanship is presented (vocational training). Chapter 5 presents a prototype of a DMI which is able to capture the motions of the whole upper-body and translate them into music following various sonification strategies. In Chapter 6, a number of other past scientific contributions (3D pose estimation, a Silent Speech Interface (SSI) etc.) or ongoing work (ergonomy of operators, egocentric computer vision for action recognition in Human-Robot Collaboration (HRC) etc.) are presented, together with a summary of my achievements, my perspectives and what I regard as my current and future research challenges.

Chapter 2

Modelling the spatiotemporal dynamics of the human body for human action recognition and forecasting

Contents

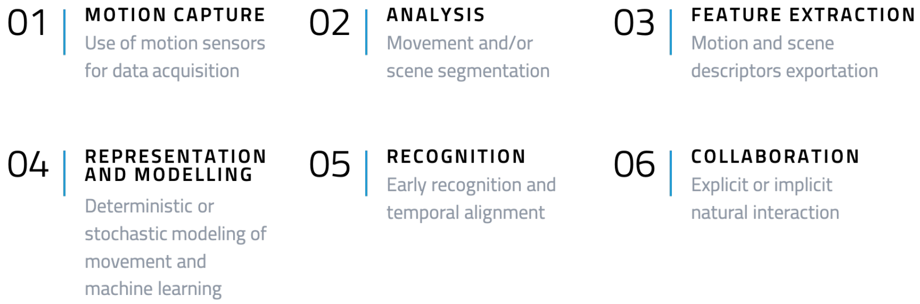
2.1 The Big Picture

The main contribution to Movement-based Human Machine Collaboration has been the development of a HAI-based perception layer, external to the machines. This layer is able to recognize human gestures early and continuously and communicate the recognition results to the machines in order for them to adapt their behaviour in accordance with human behaviour. This chapter proposes an in-depth presentation of the two fundamental steps, which are: *representation and modelling* (step 4) and *recognition* (step 5), of the whole six-step generic methodology that has been built for testing cross-application scientific questions (Figure 2.1). The first three steps and the sixth step depend on the application, the requirements and the machine. Thus, they will be presented in the chapters that follow.

The six steps of the methodology start with *motion capturing* where two categories of motion sensors are used: the wearables and the vision-based motion sensors. When computer vision is used, the scene is captured as a sequence of images and signal *analysis* is applied in order to segment the foreground from the background. Then, when the *feature extraction* step takes place, mostly the motion descriptors, but also the object descriptors, are exported. Thereafter, the *representation and modelling* step implies the static or dynamic representation of the phenomenon of gestural evolution over time. The *recognition* of motion patterns can be executed either in offline mode, where it can test the accuracy of the method or analyze the human performance, or in online mode, to facilitate *collaboration* with the machine. Thus, collaboration with machines is a natural interaction with machines, whether it is explicit, as commands, or implicit.

In this chapter, the main focus will be on the steps of representation and modelling as well as on recognition through the presentation of the Gesture Operational Model (GOM). It makes use of biomechanical principles to describe how body parts cooperate to perform a situated professional gesture, while taking into account the stochasticity of the human movement. The model is built upon several assumptions that determine both the dynamic relationship between the body entities (biomechanics) and their evolution in time (stochastics) when executing a gesture. It is based on SS representation, which provides us with a simultaneous equation system for all the body entities that are composed of a set of first-order differential equations. The coefficients of the equation system are estimated using the Maximum Likelihood Estimation (MLE), and its dynamic simulation generates a dynamic tolerance of the spatial variance of the movement over time. The scientific evidence of the GOM is evaluated through its ability to recognize gestures in a motion-time series that is modelled using continuous Hidden Markov Models (HMMs). Moreover, the system can be simulated through the solution of its equations and its forecasting ability is evaluated by comparing the similarity between the real and simulated motion data.

The scientific contribution presented in this chapter has been optimised by supervising the research engineers Gavriela Senteri and Dimitris Makrygiannis and has received funding from the Collaborate and Mingei H2020 projects.



01 MOTION CAPTURE
Use of motion sensors for data acquisition

02 ANALYSIS
Movement and/or scene segmentation

03 FEATURE EXTRACTION
Motion and scene descriptors exportation

04 REPRESENTATION AND MODELLING
Deterministic or stochastic modeling of movement and machine learning

05 RECOGNITION
Early recognition and temporal alignment

06 COLLABORATION
Explicit or implicit natural interaction

Figure 2.1: Generic methodology for gesture recognition and collaboration

The scientific evidence for the proposed methodology was tested on four industrial datasets that contain gestures and actions from a TV assembly line, from the glassblowing industry, the gestural commands to Automated Guided Vehicles (AGVs) as well as HRC in automotive assembly lines (more details in Chapter 3). The hybrid approach, with SS and HMMs, outperforms standard HMMs and a 3D CNN-based end-to-end Deep Learning (DL) architecture.

2.2 State-of-the-Art

Biomechanical, statistical, or hybrid models can describe parameters of the coordinated mechanical interaction between bones, muscles, and joints within the musculoskeletal system. Each of these models put forward a different aspect of human motion; therefore, although biomechanical modelling operates on a frame-by-frame level, whether from a kinematic or kinetic perspective, stochastics models the temporal dynamics of systems that evolve over time.

2.2.1 Biomechanical modelling

Human movement is caused by internal and external action forces that are most efficiently described by biomechanical models. These models represent the human body as a set of articulated links in a kinetic chain where movement and forces in combination are calculated by using anthropometric and postural motion data [Lu & Chang 2012]. This motion data is usually provided by Inertial Measurement Units (IMUs) for kinematics (e.g. accelerations, velocities etc.) and force sensors for kinetics (e.g. ground reaction forces from force plates). It is used as input for biomechanical models [Muller *et al.* 2020], which quantify the mechanics of the musculoskeletal system during the execution of a motor task.

A number of studies use biomechanical modelling to extract the kinematic and kinetic contributions of the joints and investigate their mechanical loading and response to ergonomic interventions. To analyze the ergonomic impact of different postures on the human joints, Newton-Euler algorithms are applied for the computation of upper body joint torques [Menychtas *et al.* 2020]. The normalized integral of joint angles and joint torques are then calculated to find the kinematic and kinetic contribution for each posture. Variations of the above-described methods are found in the spanned inverse dynamics models [Faber *et al.* 2016, Shojaei *et al.* 2016] and musculoskeletal finite element-based models [Gholipour & Arjmand 2016].

The biomechanical analysis and modelling of human movement is often approached as a multi-body kinematic optimisation problem for ergonomic, clinical [Duprey *et al.* 2017], sports or physical rehabilitation purposes and is combined with Newtonian mechanics [Zatsiorsky 2008].

2.2.2 Stochastic representation, modelling and recognition

Human motion is a stochastic process with high uncertainty. Stochastics is used to mathematically describe time-varying random processes by providing a reasoning over time through the use of internal states. Thus, it is an applied-on-time series of motion data.

HMMs are widely used in various gesture-related domains, such as in gesture recognition [Lee & Kim 1999, Mitra & Acharya 2007, Bevilacqua *et al.* 2009] or in movement generation [Calinon *et al.* 2011, Tilmanne 2013]. They are based on Markov chains and they assume that a hidden state sequence causes the observed sequence following transition principles [Rabiner 1989]. In gesture recognition, the Gesture Follower (GF) algorithm allows for continuous gesture recognition and, in turn, between the reference gesture and the incoming gesture. It can learn a gesture from a single example (one-shot learning) by associating each reference gesture to a single 'state' in the Markov chain. Time alignment occurs between the reference and the incoming gesture using Dynamic Time Warping (DTW), which also offers an estimation of the time progression of the gesture in real-time. GF recognizes *which* gesture or action is currently performed, but not *how* (e.g. expressively) it is performed.

Although the straightforward implementation of HMMs allows for state transition, they do not fully support explicit transitions between movement segments (or/alternatively primitives), e.g. for 're-initializing' the recognition to an initial state when the gesture is completed. To overcome this, Hierarchical-HMMs are implemented [Françoise 2015] and are articulated around two levels: one for segmentation and one for recognition. In a way similar to GF, it adopts a one-shot learning approach. Hierarchical-HMMs are trained with a single pre-segmented gesture, which is manually annotated. Each segment is then associated with a high-level segment state which generates the sub-models of the low-level (signal), which encodes the temporal evolution of the segment. Nevertheless, Hierarchical-HMMs still only recognize *which* gesture or action is being performed, rather than *how* it is performed.

State-Space (SS) modelling is widely used in control engineering (i.e. using the Kalman filter) to mathematically model dynamic systems as a set of input, output and state variables related by first-order differential equations. In gesture recognition, SS is implemented in the Gesture Variation Follower (GVF) [Caramiaux *et al.* 2015], where speed, scaling and rotation of the gesture are considered as state variables. Particle Filtering is then used for gesture recognition by updating the parameters of the SS models and extracting the likeliest template of the input gesture, taking into consideration the varying motion characteristics. GVF recognizes not only *which* gesture or action is performed, but also *how* it is performed.

In all cases, whether GF, GVF, or Hierarchical-HMMs, the degree of 'generalisation' of the reference gesture (tolerance in GF, GVF and variance offset in Hierarchical-HMMs) is predefined by the user. If this value is low, the algorithm will be more precise in the recognition. If it is set high, the algorithm will be less reliable, due to the fact that the model will be too general and it will lead to overlaps between classes. However, the main drawback is that the value of this parameter remains fixed during the recognition. This leads to the possibility that the system might fail to recognize some variations within the gesture, because a higher or lower value may be required for these particular parameters. Moreover, there is an impact on the time alignment between the reference gesture and the incoming gesture, which can vary significantly.

2.2.3 Deep learning for pose estimation and action recognition

Deep learning architectures for human action recognition can be categorised into two main strategies: *pose or skeleton*-based and *appearance*-based.

2.2.3.1 Pose-based action recognition

In this approach, two layers of DL are applied. The first layer consists of estimating the human pose from a sequence of RGB-D images and extracting motion descriptors, which in most cases are positions or rotations. RGB-D cameras, such as the Microsoft KINECT or the LeapMotion sensor, capture the human motion and provide data streams to DL algorithms. Such algorithms are Openpose [Cao *et al.* 2019], Alphapose [Fang *et al.* 2017] and Densepose [Güler *et al.* 2018] and perform a 2D or 3D pose estimation to extract the positions of body (visual) joints that do not necessarily correspond to the physical joints. Recovering a 3D pose from RGB images is considered more difficult than for 2D pose estimation. A number of challenges which need to be addressed and overcome in 3D pose estimation are lighting conditions, body occlusions, skin colour, clothing or overloaded backgrounds.

Recurrent Neural Network (RNN)s can model the long-term temporal correlation of the features for each body part [Shahroudy *et al.* 2016]. In another study, where hand gestures are used to virtually interact with objects, a 2-stage architecture is proposed. First, an Ego-Hand Mask Encoder Network is used for extracting the feature maps. Then, the RNN temporally discerns the discriminating features in RGB image sequences [Chalasani *et al.* 2018].

Spatial-Temporal Graph Convolutional Networks are applied in full body or hand gesture recognition, using hand skeletons as input [Yan *et al.* 2018, Li *et al.* 2019]. They learn both the spatial and temporal patterns from the motion data. Convolutional Neural Network (CNN)s have proved their efficiency in pose-based action recognition. Parallel convolutions are used for recognizing between 14 - 28 hand gesture classes, with an accuracy varying between 84% - 91% [Devineau *et al.* 2018].

2.2.3.2 Appearance-based action recognition

The second layer of DL is related to the visual cues, such as colour and edges, which are considered during the training phase of appearance-based action recognition. Various DL architectures can be considered for action recognition, such as those with the use of 3D CNNs [Tran *et al.* 2015b], 2-stream fusion networks, which are usually based on RGB and optical flow [Feichtenhofer *et al.* 2016], or 2-stream fusion networks with a 3D ConvNet for action recognition [Asadi-Aghbolaghi *et al.* 2017].

More recently, a 2-stream Inflated 3D ConvNet based on 2D ConvNet inflation has been introduced [Carreira & Zisserman 2017], which has the advantage that it can learn seamless spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and parameters.

There have been cases where pose-based and appearance-based recognition have been considered simultaneously as a multi-stream network. Vision-based and sensor-based motion data is used in [Song *et al.* 2016]. They extended a multi-stream CNN to learn spatial and temporal features from egocentric videos, as well as features from various sensors, such as an accelerometer, gyroscope etc.

Finally, an egocentric gesture recognition that combines CNNs spatiotemporal transformer modules to address issues of spontaneous head movements, while the camera is in motion, is presented in [Cao *et al.* 2017]. The challenge in egocentric computer vision is that both the background and the human body are moving simultaneously, while the camera follows the motion of the head, which is not always aligned to that of the rest of the body. A spatio-temporal transformer module transforms the 3D feature maps to a canonical view in both spatial and temporal dimensions.

2.2.4 Motion trajectory forecasting and intention prediction

The three main strategies for human motion trajectories forecasting and intention prediction use *physical-based models*, *pattern-based models* and *planning-based models*. Physical-based models are dynamic models that are explicitly defined and follow Newton's Law of Motion. Pattern-based models learn statistical behavioural patterns that emerge on the observed motion trajectories. Planning-based models follow reasoning about the intention behind the movement and the goal of the individual.

2.2.4.1 Physical-based models

Physical-based models are kinematic models that evolve over time and describe the particular human motion. They are 'time-domain' models that forecast future states of a dynamic system (human body), usually following an SS representation. The dynamic system evolves over time and parameters such as position, orientation, velocity, or acceleration are the state variables. The motion is forecast through the system's static or dynamic simulation by solving a number of equations. Constant velocity [Fardi *et al.* 2005], acceleration [Binelli *et al.* 2005], and coordinated turn [Schneider & Gavrila 2013] are commonly-used kinematic models for short-term forecasting of human motion with small uncertainty.

Kalman Filters are SS models that are used to forecast the positions of pedestrians based on kinematic models [Barth & Franke 2008, Binelli *et al.* 2005], which use velocity or acceleration as state variables. In another application, kinematic models are used to forecast the trajectories of cyclists by taking into consideration the driving force and resisting force together with its acceleration resistance, rolling resistance and air resistance components, all as state variables [Zernetsch *et al.* 2016]. In order for the states of the system to be estimated, a curve-fitting approach with motion profiles of cyclists is used, recorded by video cameras and laser scanners at a public intersection.

In 'time-domain' problems, the domain can be described by a single model or by multiple physical-based models. Such multi-models are used for forecasting human motion with a high uncertainty. Different modes of motion (e.g. sudden accelerations, linear movements, manoeuvres) represent complex behaviours (e.g. pedestrian or vehicles in public areas), using a different dynamic model for each motion mode [Kooij *et al.* 2019]. Such an approach is applied to predicting the motion of cyclists, taking into consideration their motion strategies (e.g. go straight, turn left or right at an angle of 45° or 90°) [Pool *et al.* 2017].

Physical-based approaches are appropriate when an explicit transition function can be defined for modelling the motion dynamics through endogenous and exogenous input variables. Nevertheless, they do not perform well for very complex situations (e.g. public areas with multiple agents) and they are usually used only for short-term forecasting.

2.2.4.2 Pattern-based models

Pattern-based strategies follow model-fitting approaches. Gaussian process dynamic models are used to predict pedestrian trajectories [Mínguez *et al.* 2018]. Key points on the pedestrian bodies are extracted and their 3D time-related information is reduced into two observations that are also only used for prediction. The most similar model from the multiple models of four activity types (e.g. walking, stopping, starting, and standing) is then selected to estimate future pedestrian states. In cases where the goal is to predict the motion of multiple persons, then mixtures of Gaussian processes are used to model multiple distributions for speed and joint orientation, whose flow is mapped in the prediction [Kucner *et al.* 2017].

Neural Networks are also used when time series are given as input [Sun *et al.* 2018, Xue *et al.* 2018, Srikanth *et al.* 2019]. More precisely, Long Short-Term Memory (LSTM) networks have proven the value of their performance in predicting human [Sun *et al.* 2018, Xue *et al.* 2018] and vehicle motion

[Srikanth *et al.* 2019]. For pedestrians, 2D position and orientations are usually provided to one or more LSTMs (for example when information about the scene is also taken into consideration) in order to learn human behavioural patterns from different environments. For vehicles, a simple Encoder-Decoder model is connected to a convolutional LSTM to learn vehicle temporal dynamics, including semantic images, depth information, and other vehicles' positions.

Pattern-based approaches can outperform physical-based models when a large amount of data is available and unknown dynamics are involved in the process. Both approaches can give interesting and robust results when context information is provided through observations, such as the shape and structure of the environment, external forces that the person or object is exposed to, or information about their interaction with other agents (e.g. people, vehicles, or robots) [Rudenko *et al.* 2020].

2.2.4.3 Planning-based models

Planning-based models assume rationality on the part of humans and their long-term motion goals. This approach computes path hypotheses that allow the agent to reach their motion goals by considering the impact of current actions on future motions. A predefined cost function takes into consideration the motion intention while an inferred cost function takes into consideration the observed trajectories. Multimodal hypotheses can be introduced to predict the trajectories and intended goals of pedestrians using a Bayesian framework [Best & Fitch 2015]. Another strategy is to approach the prediction as an optimization problem [Lee *et al.* 2017] by proposing a deep stochastic RNN and an Encoder-Decoder framework for trajectory prediction of multiple vehicles in complex scenes. Diverse hypothetical trajectories are considered for the agent interactions and scene semantics through a reward function. The model captures the past trajectories and incorporates the information into the inference process to improve the prediction accuracy.

Planning-based approaches consider an explicit definition of the goals of the humans and other scene agents. They usually perform better for long-term predictions than physical-based approaches and they are better able to manage generalization challenges than pattern-based approaches. Nevertheless, the more the complexity of the prediction increases (e.g. long-term predictions, multiple agents, size of the environment), the more the training becomes heavy.

2.3 Industrial datasets and pose estimation

Four industrial real-life datasets $GV_{i\in[1,4]}$ were collected from a household appliances manufacturer, an AGV manufacturer, a glassblowing workshop and an automotive industry and these are presented in Figure 2.2.

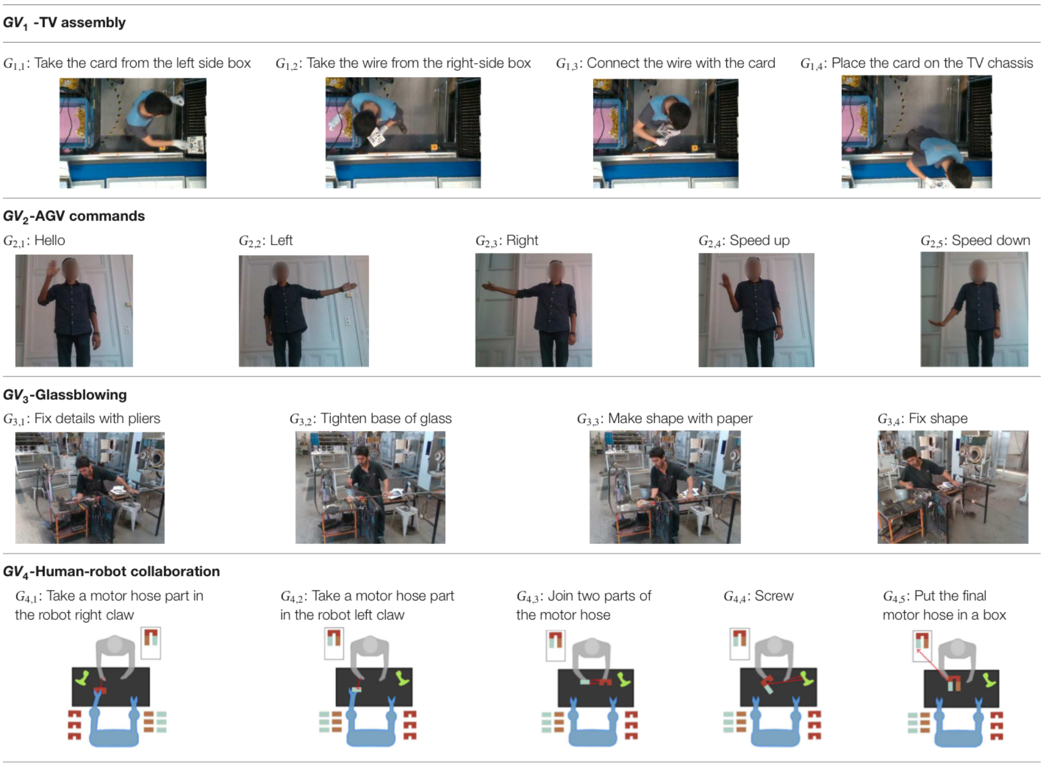

GV1 -TV assembly
G1,1: Take the card from the left side box
G1,2: Take the wire from the right-side box
G1,3: Connect the wire with the card
G1,4: Place the card on the TV chassis
GV2-AGV commands
G2,1: Hello
G2,2: Left
G2,3: Right
G2,4: Speed up
G2,5: Speed down
GV3-Glassblowing
G3,1: Fix details with pliers
G3,2: Tighten base of glass
G3,3: Make shape with paper
G3,4: Fix shape
GV4-Human-robot collaboration
G4,1: Take a motor hose part in the robot right claw
G4,2: Take a motor hose part in the robot left claw
G4,3: Join two parts of the motor hose
G4,4: Screw
G4,5: Put the final motor hose in a box

Figure 2.2: Gesture vocabularies of TV assembly, AGV commands, Glassblowing and Human-robot collaboration

GV_1 was recorded from within the household appliance factory of Arçelik in Turkey. It includes four gestures where the operator takes the electronic card from one box and then takes a wire from another, connects them, and places them on the TV chassis. Although all the gestures are performed by a single user, there are a lot of different positions of the operator within the workspace for each gesture. This dataset is noisy and it provided an opportunity to examine the performance of both the pose estimation and the recognition algorithm with such data.

GV_2 is a joint dataset collection between engineers and operators of ASTI in Spain and researchers from the Centre for Robotics at MINES ParisTech. It includes gestural commands for controlling an AGV, where the operator initiates the communication with the AGV by shaking the palm (waving) and gives a command for turning to the left or to the right by raising her respective arms. The user can also speed up the AGV by raising the right hand three times or reduce speed by rolling the right hand away from the hips. It is a multi-user dataset that contains gestures that could also be considered to be ‘everyday gestures’.

GV_3 was recorded at the European Centre for Research and Training in the Glassworks of CERFAV, in France. It contains four gestures performed by a glassblower when creating a water carafe. The craftsman puts the pipe on the metallic structure and performs various manipulations of the glass by using tools, such as pliers. He starts by shaping the neck of the carafe with the use of pliers, then he tightens the neck to define the transition between the neck and the curved vessel. In his right hand, he holds a specific paper and shapes the curves of the blown part and finalizes the object and fixes the details by using a metallic stick. It is a mono-user dataset that involves a high level of dexterity.

Finally, GV_4 is related to a real-life human–robot collaboration scenario that was recorded in the automotive assembly lines of PSA Peugeot Citroën (PSA Group). It is presented in detail in Section 3.6.

For GV_1, GV_2 and GV_3, each image sequence is imported to the OpenPose framework, which detects body keypoints on the RGB image and extracts a skeletal model together with the 2D positions of each body joint [Cao *et al.* 2018] (Figure 2.3). These joints are not necessarily physical joints. In most cases they correspond to physical joint centres and the coordinates of each joint are derived by the width and height of the camera. For GV_4, 3D hand positions are extracted from top-mounted depth imaging by detecting keypoints on the depth map, as is described in Section 3.6.

2.4 The Gesture Operational Model

When a skilled individual performs a professional situated gesture, the whole body is involved, thus combining theoretical knowledge with practical motor skills. Effective and accompanying body movements are harmonically coordinated to execute a given action. The expertise in the execution of professional gestures is characterized by precision and repeatability, while the body is continuously shifting from one phase to another, for example, from specific postures (small tolerance for spatial variance) to ample movements (high tolerance for spatial variance). For each phase of the movement, each body entity, for example, articulation or segment, moves in a multidimensional space over time. When considering the 2D motion descriptors of the movement, two mutually dependent variables represent the entity, for example, X and Y positions. Each of these variables is associated with the other, thus creating a bidirectional relationship between them. Furthermore, they also depend on their history, whereas some entities might 'work together' to execute an effective gesture, for example, when an operator assembles two parts. However, a unidirectional dependency might be observed when one entity influences the other entity and not vice versa, as well as a bidirectional dependency when both entities influence each other, for example, when a potter shapes the clay with both hands.

The above observations on situated body movements can be translated into a functional model, which we define here as the GOM, which describes how the body or skeletal entities of a skilled individual are organized to deliver a specific result (Figure 2.3). It can be hypothesised that each of the assumptions of 'intrajoint association', 'transitioning', 'intralimb synergies' and 'intralimb mediation' contribute at a certain level to the production of a gesture. As far as intralimb mediation is concerned, it can be broken down to 'interjoint serial mediation' and 'interjoint non-serial mediation' assumptions. The proposed model works perfectly for all three dimensions (X, Y, and Z), but for reasons of simplicity, it will here be presented only for the two dimensions of X and Y. In addition, in this work, only positions are used, but the model is designed to be able to receive joint angles as input as well.

H1: Intra-joint association

It is hypothesized that the movement of each body part (e.g. the right hand) is described through positions from the Cartesian coordinate system. Consequently, the motion trajectory of each body part P is broken down into x and y coordinates, which generate two mutually dependent variables: P_x and P_y. It is assumed that there is a bidirectional relationship between P_x and P_y that is defined as an *intra-joint assumption.*

H2: Inter-limb synergies

It is assumed that some body parts work together to achieve certain motion trajectories, e.g. P and P', which is defined here as *inter-limb synergies.* For example, when hands cooperate to assemble two parts.

H3.1: Inter-joint serial mediation

It is assumed that a body part P may depend on a neighboring part P'' to which it is directly connected. In cases where this assumption is statistically significant, there is an *inter-joint serial mediation.* For example, a glassblower, while using the pipe, moves his/her wrists along with his/her shoulders and elbows.

H3.2: Inter-joint non-serial mediation

It is assumed that the movement of a body part P depends also on a non-neighbouring part P''' of the same limb; for example, the movement of the wrist may depend on the movement of the elbow and shoulder. In cases where this assumption is statistically significant, there is an *inter-joint non-serial mediation*. Thus, it is highly likely that both direct and indirect dependencies simultaneously occur in the same gesture.

H4: Transitioning
It is also assumed that each variable depends on its own history, also called inertia effect. This means that the current value of each variable depends on the values of previous times, also called lag or dynamic effect, which is defined here as *transitioning*.

Thus, an example of the representation of those assumptions for the x-axis of the body part P with an inertia effect of one previous time would be as follows:

$$P_x(t) = P_y(t-1) + P'_x(t-1) + P''_x(t-1) + P'''_x(t-1) + P_x(t-1) \tag{2.1}$$

where, $P_x(t)$ is the variable of the x coordinate of P that is to be estimated, $P_y(t-1)$ is the variable for the movement of P on y-axis in the previous time stamp and is linked with the dependency between the movement on y at $t-1$ and on x at t (H1: *intra-joint assumption*), $P'_x(t-1)$ is the variable on x for the part P', which is on a different limb from P', and is linked with the dependency between the movement of P at t and of P' at $t-1$ (H2: *inter-limb synergies*), P'' is a neighbour of P while P''' is not and all three belong to the same limb and both variables $P''_x(t-1)$ and $P'''_x(t-1)$ are linked with the dependencies between the movement of P at t and of P'' at $t-1$ (H3.1: *inter-joint serial mediation*) as well as of P at t and of P''' at $t-1$ (H3.2: *inter-joint non-serial mediation*). Finally, the variable $P_x(t-1)$ is linked with the dependency between the movement of P at t and $t-1$ (H4: *transitioning*).

The Equation (2.1) describes a first-order auto-regressive model, whose order generally depends on the data characteristics and the specificities of the experiment. By using a second-order auto-regressive model a better forecasting ability may be obtained. Thus, when the transitioning of the GOM depends on two previous times, the Equation (2.1) is re-written as follows:

$$P_x(t) = P_y(t-1) + P'_x(t-1) + P''_x(t-1) + P'''_x(t-1) + P_x(t-1) + P_x(t-2) \tag{2.2}$$

2.5 State-Space Representation

N first-order differential equations are generated by concatenating the dynamics of the GOM, which is an N-order system, where $N =$ *space dimension* ×

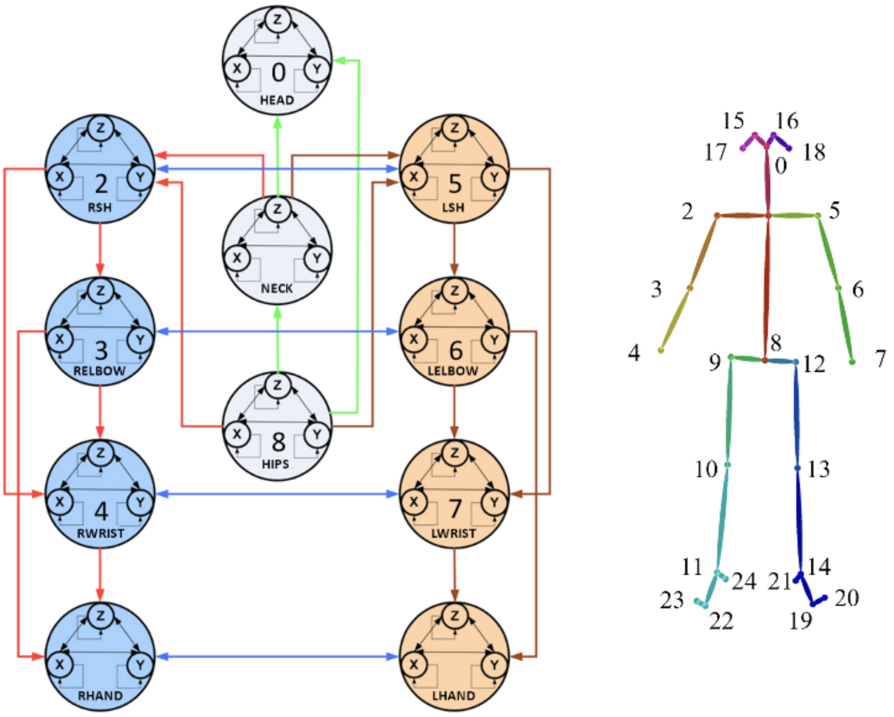


Figure 2.3: The Gesture Operational Model for the upper-body part in 3D space with its four assumptions: H1: Intra-joint association (black arrows), H2: Inter-limb synergies (blue arrows), H3.1 Intra-limb serial mediation and H3.2 Intra-limb non-serial mediation (green arrows for torso, red for right arm and dark red arrows for left arm) and H4: Transitioning (dotted arrows). The numbers on the GOM correspond to the joint representation from the OpenPose framework (left skeleton).

number of body parts. Finally, the steps which are followed are the estimation of the model, including the verification of its structure and of its forecasting and simulation ability.

According to the theory of State-Space (SS) modelling, there is the possibility of the coefficients dynamically changing over time. With SS modelling, the way the dynamic system is changing is a function of its current state. For example, in GOM, the way the x position of a body part P is changing, is a function of the previous positions of P at x; thus $P_x(t-1)$ is a lagged endogenous variable. Moreover, there are a number of exogenous variables that also influence the system, which in the case of GOM are the previous position of P at y, but also of the previous positions on the same axis of a part P' (different limb) as well as of its neighbour P'' and non-neighbour P''' parts (same limb). The endogenous variables constitute for it the minimum set of variables that fully describe it, which means that there is enough information about the changing part of the gesture and that its future behaviour can be forecast.

An SS model consists of: a. one *measurement or observation equation* relating an $n-$dimensional vector of observed variables $y(t)$ (output), to an $m-$dimensional vector of unobserved endogenous variables $s(t)$ (Markovian states), given the excitation input $u(t)$ and b. one *transition or state equation* that describes the evolution of the state vector over time. The observation equation is the signal through which the hidden state is observed and it shows the relationship between the system's state, the excitation input from exogenous parameters, and the output signal. Therefore, SS is defined as follows:

$$y(t) = Cs(t) + Du(t) \tag{2.3}$$

$$s(t) = As(t-1) + w(t) \tag{2.4}$$

where Equation (2.3) is the *observation equation* and Equation (2.4) is the *transition equation.* In (2.3), C is the output matrix that describes how the states are combined to get the output signal and D is the feed-through matrix that is used to allow the exogenous variables to bypass the system altogether and feed-forward to the output. In (2.4), A is the transition matrix that describes how all the internal states are connected to each other and underline the dynamics of the system and $w(t)$ is a white noise vector. Based on a number of experiments, it was observed that Gaussian disturbances (white noise) from the signal of the motion sensors do not affect the estimation results, and thus will not be taken into consideration.

In practice, the output matrix C is equal to [1 1] indicating that all the lagged endogenous variables are kept, while the feed-through matrix D consists of the coefficients for the exogenous variables of the input vector $u(t)$. Thus, Equations (2.3) and (2.4) are re-written as follows:

$$s(t) = A * s(t-1) = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} \begin{bmatrix} P_x(t-1) \\ -P_x(t-2) \end{bmatrix} = \begin{bmatrix} a_1 P_x(t-1) \\ -a_2 P_x(t-2) \end{bmatrix} \tag{2.5}$$

$$\begin{aligned} P_x(t) &= \begin{bmatrix}1 & 1\end{bmatrix} s(t) + \begin{bmatrix}a_3 & a_4 & a_5 & a_6\end{bmatrix} \begin{bmatrix} P_y(t-1) \\ P'_x(t-1) \\ P''_x(t-1) \\ P'''_x(t-1) \end{bmatrix} \Rightarrow \\ P_x(t) &= \begin{bmatrix}1 & 1\end{bmatrix} \begin{bmatrix} a_1 P_x(t-1) \\ -a_2 P_x(t-2) \end{bmatrix} + \begin{bmatrix}a_3 & a_4 & a_5 & a_6\end{bmatrix} \begin{bmatrix} P_y(t-1) \\ P'_x(t-1) \\ P''_x(t-1) \\ P'''_x(t-1) \end{bmatrix} \end{aligned} \tag{2.6}$$

By solving the Equations (2.5) and (2.6), we obtained the following general SS representation for the part P:

$$\begin{aligned} P_x(t) = &\underbrace{a_1 P_x(t-1) - a_2 P_x(t-2)}_{\text{H4}} + \\ &+ \underbrace{a_3 P_y(t-1)}_{\text{H1}} + \underbrace{a_4 P'_x(t-1)}_{\text{H2}} + \underbrace{a_5 P''_x(t-1)}_{\text{H3.1}} + \underbrace{a_6 P'''_x(t-1)}_{\text{H3.2}} \end{aligned} \tag{2.7}$$

where a_1 to a_6 are the coefficients that quantify the contribution of each assumption related to the gesture.

In order to compute the coefficients of the Equation (2.7), the Kalman filter was used. The Kalman filter minimizes the mean square error of the estimated parameters. It is a recursive method since new measurements are processed as they arrive and it consists of two recursive steps: the prediction and the update. Below, the two steps using the Kalman filter are presented.

Prediction

- Predicted state estimate: $\hat{s}(t \mid t-1) = A\hat{s}(t-1 \mid t-1)$
- Predicted error covariance: $P_{ec}(t \mid t-1) = AP_{ec}(t-1 \mid t-1) + A^T + G(t)$, where $G(t)$ is Gaussian disturbance.

Update

- Measurement pre-fit residual: $\tilde{y}(t) = P_x(t) - C\hat{s}(t \mid t-1)$
- Pre-fit residual covariance: $S(t) = CP_{ec}(t \mid t-1)C^T + G(t)$
- Optimal Kalman gain: $K(t) = P_{ec}(t-1)C^T S^{-1}(t)$
- Update state estimate: $\hat{S}(t \mid t) = \hat{S}(t \mid t-1) + K(t)\tilde{y}(t)$
- Update estimated covariance: $P_{ec}(t \mid t) = I - K(t)CP_{ec}(t \mid t-1)$

- Measurement post-fit residual: $\hat{y}(t \mid t) = P_x(t) - C\hat{s}(t \mid t)$

The steps described above are recursive and applied at each state until the optimal estimates are computed.

2.6 Learning with Hidden Markov Models

Hidden Markov Models follow the principles of Markov chains that describe stochastic processes. They are commonly used to model and recognize human gestures and are structured using two different types of probabilities, the transition probability from one state to another and the probability for a state to generate specific observations on the signal [Bakis 1976]. In our case, each professional gesture is associated with an HMM, whereas the intermediate phases of the gesture constitute internal states of the HMM. According to our four datasets, these gestures define one Gesture Vocabulary per dataset: $GV_{i\in[1,4]} = \{G_j\}_{j\in N}$.

Let us consider a hidden sequence of states $q = \{q_1, q_2 \ldots, q_k\}_{k\in N}$, corresponding to all the intermediate phases of a professional gesture. The transition probability a_{ij} between the states q_i and q_j is provided by the transition matrix $Q = [a_{11}, ..., a_{ij}, ..., a_{kk}]_{k\in N}$. The sequence q is supposed to generate a sequence of observation vectors $O = \{o_1, o_2 \ldots, o_k\}_{k\in N}$. We assume that the vectors o_k depend only on the state q_k. From now on, the likelihood that the observation o is generated by the state q will be defined as $\mathcal{L}(o|q)$. It is important to outline that in our modelling structure, each internal state of the model depends only on its previous state (first-order Markov property). Consequently, the set of the models for all gestures for every gesture vocabulary is $GV_{i\in[1,4]} = \{HMM_i\}_{i\in N}$, where $\{HMM_i\} = (\varrho_i, Q_i, \mathcal{L}_i)$ are the parameters of the model and ϱ_i is the initial state probability. Thus, the recognition becomes an issue of solving three specific problems: evaluation, recognition, and learning (Dymarski, 2011). Each one of those problems was solved with the use of the algorithms Viterbi [Rabiner 1989], Baum's 'forward' [Baum *et al.* 1972], and Baum–Welch, respectively [Dempster *et al.* 1977].

2.7 Gesture recognition

In the recognition phase, the main goal is to recall, with high precision, the hidden sequence of internal states q that correspond to the sequences of the observation vectors. Thus, let us consider the observation of motion data O, which need to be recognized. Every $HMM_{i\lambda\in[1,j]}$ of a given $GV_{i\in[1,4]}$ that contains j gestures generate the likelihood $\mathcal{L}(O|HMM_{i\lambda})$. If there is a $HMM_{i\lambda\in[1,j]}$ that generates $\mathcal{L} \geq 0.55$, then it is considered that O is generated by $G_{i,\lambda}$. Otherwise, the following quantity is computed for every $SS_{i\xi}$ of GV_i (confidence control):

$$SS_{i\xi}^{score} = \frac{1}{1 + d(O, O_{i\xi}^{s})} \tag{2.8}$$

where d is the minimum distance between the simulated values $O_{i\xi}^{s}$ from the model $SS_{i\xi}$ and the original observations O. Then, for every $SS_{i\xi}$ of GV_i the likelihood $\mathcal{L}(O|HMM_{i\xi}^{SS})$ is computed as follows:

$$\mathcal{L}'(O|HMM_{i\xi}^{SS}) = \mathcal{L}(O|HMM_{i\xi}) \times SS_{i\xi}^{score} \tag{2.9}$$

and we arrive at the final formula providing the way the algorithm recognizes the observation of motion data O,

$$R_{GV_i}(O) = \begin{cases} \max_{\lambda}^{j}(\mathcal{L}(O|HMM_{i\lambda}), & \max(\mathcal{L}(O|HMM_{i\lambda})) \geq 0.55 \\ \max_{\xi}^{j}(\mathcal{L}'(O|HMM_{i\xi}^{SS}), & \max(\mathcal{L}(O|HMM_{i\lambda})) < 0.55 \end{cases} \tag{2.10}$$

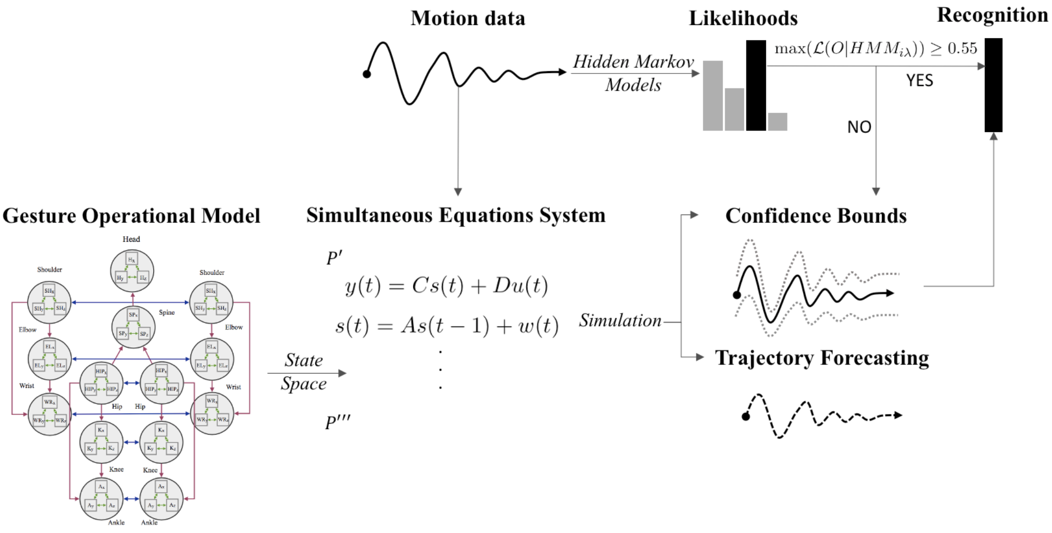


Figure 2.4: Methodological overview

2.8 Statistical significance and simulation of the models

By investigating the significance level of the coefficients for each variable (p-value), an in-depth understanding of the gesture can be obtained. More precisely, answers

can be provided to a number of questions, such as:
- *which assumptions are meaningful for this gesture?*
- *which body parts contribute more to the gesture?*
- *is a body part moving more on a single axis than in 2D/3D space ?*
- *is there any fast or slow speed of change for a body part?*

In addition, interesting uses of the coefficients of the GOM can also be considered, such as:
- *selecting the appropriate feature with the aim of recognising the gesture*
- *analysing the body dexterity of a professional expert*
- *analysing ergonomic risks related with the gesture itself etc.*

A number of examples can demonstrate the potential of GOM. In the Equation (2.11), the SS model of $G_{2,1}$ for the right wrist $RWRISTx(t)$ is presented. $G_{2,1}$ is a 'hello' waving gestural command to an AGV where the right wrist is moving across the x-axis and the left wrist remains still. No important intra-limb mediation was expected and neither H3.1 nor H3.2 is taken into consideration for the structure of this model. According to the p-values for $RWRISTx(t)$ and the coefficients of the predetermined variables, there is an intra-joint association (H1) between x and y axes ($p = 0.03 < 0.05$) but also a small value for the coefficient of $RWRISTy(t-1)$, which means that the movement over the y-axis is small. Moreover, the transitioning assumption is meaningful. Nevertheless, the very small value of the coefficient for $RWRISTx(t-2)$ could be interpreted as a possibility to use only a 1st-order autoregressive model. The assumption for an inter-limb synergy (H2) is not meaningful for this gesture ($p = 0.44 > 0.05$) and it confirms the initial observation where the right wrist is moving while the left is not. Leveraging on the analysis of the various p-values and coefficients from the Equation (2.11), there is not any important exogenous impact (e.g. from body parts other than the right wrist) on the way the individual is performing the specific gestural command, which means that it is more a 'routine' rather than a 'dexterous' gesture.

$$\begin{aligned} RWRISTx(t) = {} & \underbrace{(-0.06)\, RWRIST_y(t-1)}_{\text{H1: p = 0.03}} + \underbrace{(1.34)\, RWRIST_x(t-1)}_{\text{H4: p = 0.00}} - \\ & - \underbrace{(-0.04)\, RWRIST_x(t-2)}_{\text{H4: p = 0.00}} + \underbrace{(-0.67)\, LWRIST_x(t-1)}_{\text{H2: p = 0.44}} \end{aligned} \tag{2.11}$$

Another example of a GOM is the 'tighten the base of glass' gesture $G_{3,2}$ from the glass-blowing vocabulary. In $G_{3,2}$, the expert glass-blower first uses the left arm to rotate the pipe and afterwards he shapes the transition between the neck and the curved vessel with his right hand. The left wrist is performing rotational movements, exclusively over the y-axis (on a plane perpendicular to the camera), in order to revolve the pipe, but these movements also involve his left shoulder, which is confirmed by the $p = 0.00 < 0.05$ for $LSHy(t-1)$ (Equation 2.13). The right wrist is performing micro-movements in order to adjust the transition part; thus, his right shoulder remains still, which is also confirmed by the $p = 0.56 > 0.05$ for $RSHy(t-1)$. As a result, the non-serial mediation assumption (H3.2) is not meaningful for $RWRISTy(t)$ (Equation 2.12). Nevertheless, in the Equation 2.13, a $p = 0.12 > 0.05$ is obtained for the $LWRISTy(t-2)$, which means that a 1st-order auto-regressive model can be used for $LWRISTy(t)$. The right wrist $RWRISTy(t)$ is moving synergistically with the left wrist $LWRISTy(t-1)$, which initially defines the exact position of the whole neck to permit the $RWRISTy(t)$ to shape the object. This unidirectional relationship is confirmed by the fact that the intra-limb synergy (H2) is meaningful for $RWRISTy(t)$, but not for $LWRISTy(t)$. Furthermore, the intra-joint association (H1) is meaningful for $RWRISTy(t)$, while it is not for $LWRISTy(t)$, because the left wrist is not moving over the x-axis.

$$\begin{aligned} RWRISTy(t) = {} & \underbrace{(0.29)\, RSH_y(t-1)}_{\text{H3.2: p = 0.56}} + \underbrace{(0.37)\, RELBOW_y(t-1)}_{\text{H3.1: p = 0.00}} + \\ & + \underbrace{(-1.09)\, RWRIST_x(t-1)}_{\text{H1: p = 0.00}} + \underbrace{(-0.16)\, LWRIST_y(t-1)}_{\text{H2: p = 0.04}} + \\ & + \underbrace{(1.13)\, RWRIST_y(t-1)}_{\text{H4: p = 0.00}} - \underbrace{(-0.17)\, RWRIST_y(t-2)}_{\text{H4: p = 0.00}} \end{aligned} \tag{2.12}$$

$$\begin{aligned} LWRISTy(t) = {} & \underbrace{(0.93)\, LSH_y(t-1)}_{\text{H3.2: p = 0.00}} + \underbrace{(0.41)\, LELBOW_y(t-1)}_{\text{H3.1: p = 0.01}} + \\ & + \underbrace{(0.11)\, LWRIST_x(t-1)}_{\text{H1: p = 0.27}} + \underbrace{(0.01)\, RWRIST_y(t-1)}_{\text{H2: p = 0.36}} + \\ & + \underbrace{(1.05)\, LWRIST_y(t-1)}_{\text{H4: p = 0.00}} - \underbrace{(-0.11)\, LWRIST_y(t-2)}_{\text{H4: p = 0.12}} \end{aligned} \tag{2.13}$$

The simulation of the models is based on the solution of their simultaneous equations system. An example of a static forecast (one-step-ahead) is presented in Figure(2.5) where there are both the real observations and the simulated values from the SS model for the right wrist. The behaviour of the models is quite satisfactory since both curves are very close.

2.9 Classification assessment and comparison with an end-to-end 3DCNN

The standard metrics of average (or mean) precision $\overline{P}$ and precision P_i for gesture i, average recall $\overline{P}$ and recall R_i for gesture i, and F-score F, as well as the accuracy

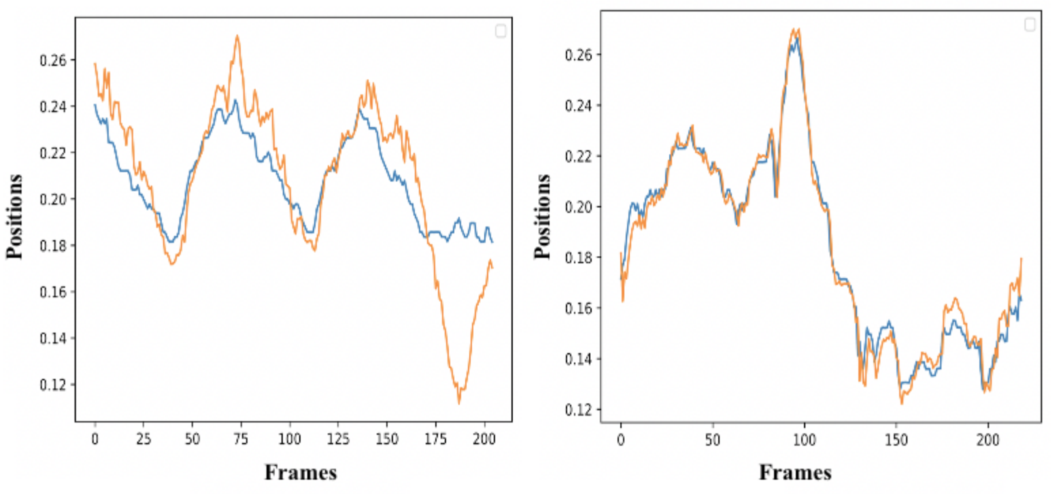


Figure 2.5: Examples of real motion observations (blue) and simulated values (orange) from the $RHANDx(t)$ SS model of the gestures $G_{3,1}$ (left) and $G_{3,4}$ (right).

were used to assess the recognition accuracy of the algorithm

$$\overline{P} = \frac{\#\ of\ True\ Max\ Likelihoods}{\#\ of\ True\ Max\ Likelihoods + \#\ of\ False\ Max\ Likelihoods}$$

where *# of True Max Likelihoods* is the number of test sequences for which the relevant model (or template) of the performed gesture gave maximum likelihood, while *# of False Max Likelihoods* is the number of test sequences where a non-relevant model gave maximum likelihood. Thus, $\overline{P}$ is the total number of relevant gestures that are correctly recognised, divided by the number of true and false positives for all the gestures, while P_i is only for gesture i.

$$\overline{R} = \frac{\#\ of\ True\ Max\ Likelihoods}{\#\ of\ True\ Max\ Likelihoods + \#\ of\ False\ Non\_Max\ Likelihoods}$$

where *# of False Non_Max Likelihoods* is the number of test sequences for which no maximum likelihood is given by the relevant model (or template) for the performed gesture. Thus, $\overline{R}$ is the total number of relevant gestures that are correctly recognised, divided by the number of true positives and false non positives for all the gestures, while R_i is only for gesture i.

$$F = 2\frac{\overline{P} \times \overline{R}}{\overline{P} + \overline{R}}$$

where F is the harmonic mean of average precision and recall and provides a global recognition performance index.

$$Accuracy = \frac{\#\ of\ Correct\ Classifications}{Total\ \#\ of\ Classifications}$$

which is the proportion of correctly recognised gestures.

In order to compare the results of this approach with other classification techniques, an end-to-end 3DCNN was used to classify the gestures of the first three vocabularies described in Industrial Datasets and Gesture Vocabularies. More precisely, a 3DCNN was initially trained on spatiotemporal features from a medium-sized UCF-101 video dataset [Soomro *et al.* 2012], and the pretrained weights were used to fine-tune the model on small-sized datasets, which included images of operators performing customized gestures in industrial environments. The architecture of the network was based on four convolution and two pooling layers, one fully connected layer and a softmax loss layer to predict action labels [Tran *et al.* 2015a]. It was trained from scratch on the UCF-101 video dataset, using a batch size of 32 clips and an Adam optimizer [Kingma & Ba 2014] for 100 epochs, within a Keras framework [Chollet *et al.* 2015]. The entire network was frozen, and only the last four layers were fine-tuned on customized datasets by means of backpropagation.

GV_1 contained 4 classes, with 44 - 48 repetitions for each. Four hidden states were used for the training of the HMMs, while a simplified GOM with only x and x positions for the two wrists were used for training and recognition. HMMs provided a recall superior to 90% in three out of four gestures of GV_1. When the confidence control was applied through the SS representation, in most cases there was a further significant improvement in the recall, such as with +15.91% between $HMM_{1,3}^{SS}$ and $HMM_{1,3}$. This improvement in the recall of $G_{1,3}$ can be justified by the various optional locations inside the workplace that the operator can have for connecting the wire with a very small card outside. The precision of $G_{1,3}$ was also positively impacted by the confidence control of the SS representation, thus improving it by +7.47%. This conclusion confirms the initial claim that combining HMMs with the confidence control of SS representation can potentially give better results for gestures where the human needs to obtain specific positions e.g. when manipulating objects. The $3DCNN$ improved the recall of $G_{1,3}$ by +2.1% compared with both the HMMs and the SS representation. However, in total, the HMM^{SS} outperformed both the HMM and the $3DCNN$, with a total accuracy and F-score of approximately 96.2% (Figure 2.6).

GV_2 contained 5 classes, with 16 repetitions of each gesture and between 1 - 11 hidden states, which optimized the performance of the models. Positions for the wrist, elbow, and shoulder joints for each arm, along with the neck, were used for training. The performance of $HMM_{2,1}$, $HMM_{2,4}$ and $HMM_{2,5}$ was optimised when they were built using an ergodic topology, since they all had a part of the gesture in common. $HMM_{2,2}$, $HMM_{2,3}$ were built using a left-to-right topology.

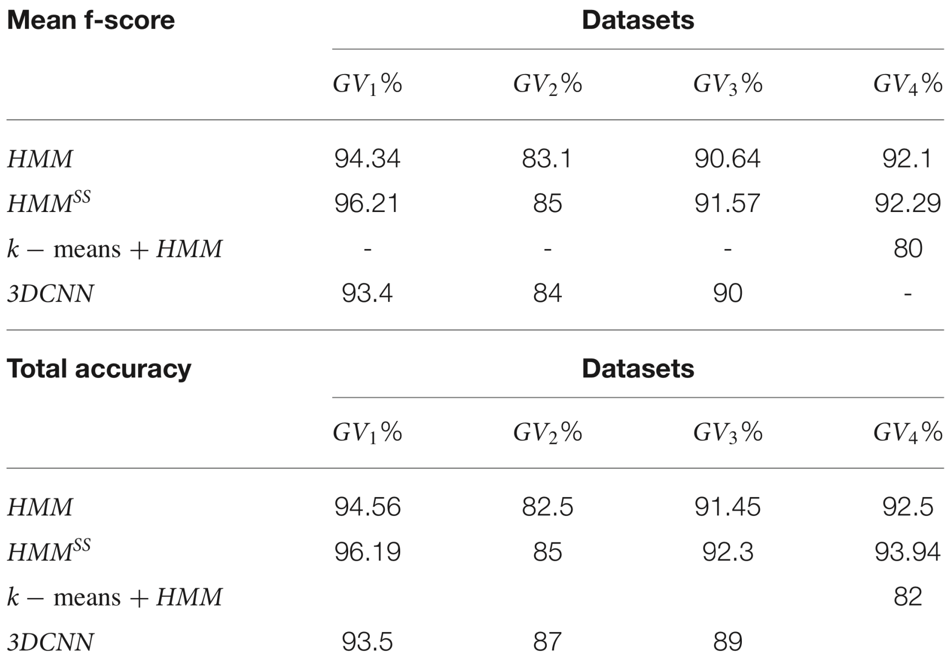
By adding the SS representation, precision in recognising all the gestures was improved, while recall was decreased for $G_{2,2}$ and $G_{2,5}$. With regard to the $3DCNN$, both precision and recall for the network was 100% for both $G_{2,2}$ and $G_{2,3}$ but for $G_{2,5}$ the precision was 66.66%, which is extremely low compared with the 100% of precision for HMM and HMM^{SS}. According to what is presented in Figure 2.6, the HMM^{SS} outperforms the $3DCNN$ with a total accuracy of approximately $+3.3\%$, while the F-score is $+1.57\%$, compared with the $3DCNN$.

GV_3 consisted of 4 different gestures, with 35, 34, 21, and 27 repetitions respectively. A significant variability in the number of hidden states (between 5 - 20), that optimised the performance of the models, was observed. Positions for the wrist, elbow, and shoulder joints for each arm, along with the neck, were used for training. When applying the confidence control of the SS representation on the HMMs, precision increased for all gestures, except for $G_{3,4}$, while the recall remained stable for all gestures except for $G_{3,3}$, which increased by 4%. HMM^{SS} outperformed both HMM models and the $3DCNN$ architecture, as shown also in Figure 2.6, with a total accuracy of approximately $+4\%$ and an F-score of $+1.5\%$, compared to the $3DCNN$.

GV_4 consisted of 5 different gestures. The models were trained with 25 clusters and 12 hidden states. HMM^{SS} outperformed continuous HMM models with $+0.19\%$ for the F-score and $+1.44\%$ for accuracy and discrete k-Means$+HMM$, with $+12.29\%$ for the F-score and $+11.94\%$ for accuracy, as presented in Figure 2.6.

2.10 Forecasting ability for motion trajectories

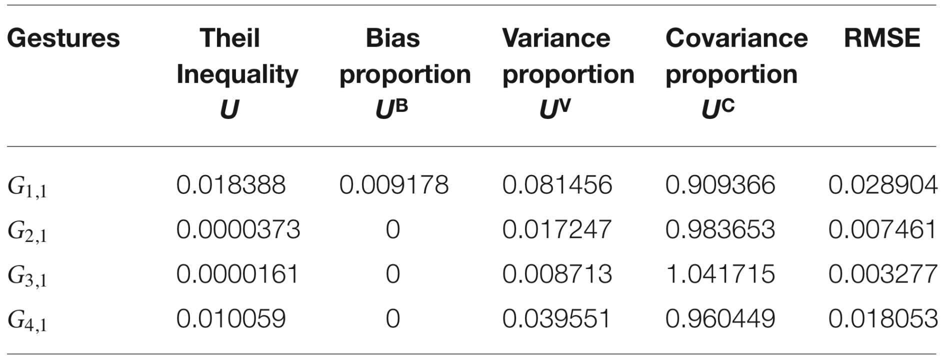
To evaluate the capability of the four SS models to provide representation of movement that can be used to explain the assumptions of the two-part GOM, Theil's inequality coefficient U was computed together with the breakdown of its components into the inequality of bias proportion U^B, variance proportion U^V and covariance proportion U^C. While U^B examines the relationship between the means of the actual values and the forecasts, U^V considers the ability of the forecast to match the variation in the actual series and U^C captures the residual unsystematic element of the forecast errors [Makridakis *et al.* 2008]. Thus, $U^B + U^V + U^C = 1$. The Theil inequality coefficient measures how close the simulated variables are to the real variables, and it is therefore able to capture values between 0 to 1. The closer to 0 the value of this factor is, the better the forecasting of the variable. Also, the forecasting ability of the model is better when U^B and U^V are close to 0 and U^C is close to 1. The computed coefficients are presented in Figure 2.7 and have resulted in a sufficiently accurate forecasting of the performance in the simulated model.



Mean f-score	**Datasets**			
	GV_1%	GV_2%	GV_3%	GV_4%
HMM	94.34	83.1	90.64	92.1
HMM^{SS}	96.21	85	91.57	92.29
$k-$ means $+$ *HMM*	-	-	-	80
3DCNN	93.4	84	90	-

Total accuracy	**Datasets**			
	GV_1%	GV_2%	GV_3%	GV_4%
HMM	94.56	82.5	91.45	92.5
HMM^{SS}	96.19	85	92.3	93.94
$k-$ means $+$ *HMM*				82
3DCNN	93.5	87	89	

Figure 2.6: Comparison of HMM, HMM^{SS} with $3DCNN$ for GV_1, GV_2 and GV_3 and with k-Means$+HMM$ for GV_4



Gestures	**Theil Inequality *U***	**Bias proportion U^B**	**Variance proportion U^V**	**Covariance proportion U^C**	**RMSE**
$G_{1,1}$	0.018388	0.009178	0.081456	0.909366	0.028904
$G_{2,1}$	0.0000373	0	0.017247	0.983653	0.007461
$G_{3,1}$	0.0000161	0	0.008713	1.041715	0.003277
$G_{4,1}$	0.010059	0	0.039551	0.960449	0.018053

Figure 2.7: Theil inequality coefficient and root mean squared error for one example of $RWRIST_xt$

Figure 2.8 presents an example of trajectory forecasting for $G_{2,4}$. More specifically, using all the SS models of GV_2, it is asked to forecast their variable $RWRISTx(t)$, when original motion values from $G_{2,4}$ are provided for initialisation. The plotting of the similarity or distance metric from the DTW is shown in Figure 2.8, taking as input for every time t: 1) the simulated values of the $RWRISTx(t)$ from the models of GV_2 when providing it with 2 real observation values, and 2) the real observations between t and the end of the sequence. The distance metric becomes minimal, resulting in high similarity, from the very beginning for the SS model of $G_{2,4}$.

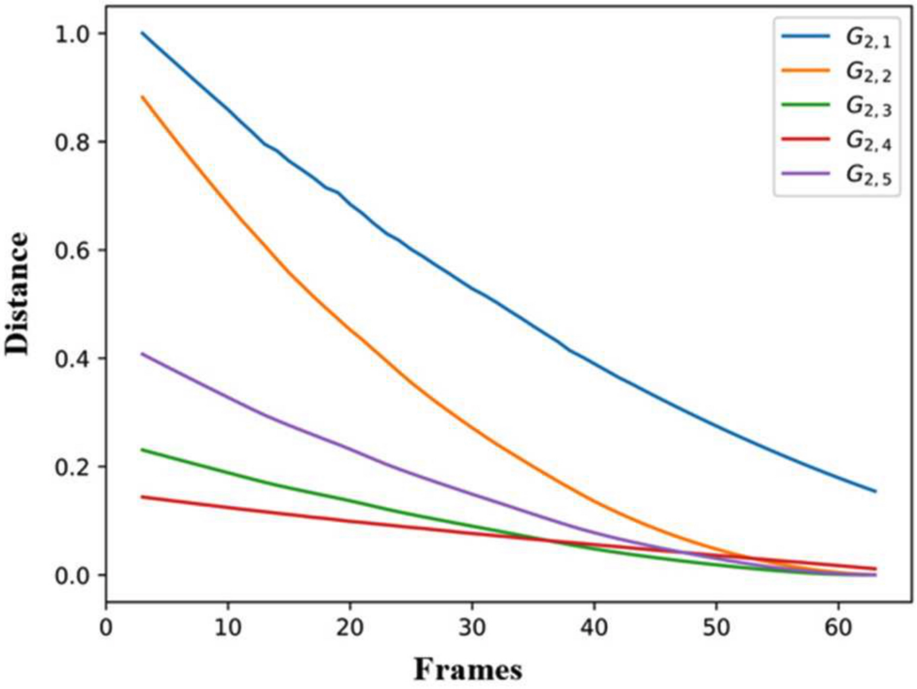


Figure 2.8: Similarity comparison on forecasted values provided by all the models of GV_2 for $RWRISTx(t)$

2.11 Sensitivity analysis

As mentioned previously, the GOM depicts all the dynamic relationships that occur during the process of the execution of a gesture. The sensitivity analysis of the simulated GOM follows two steps. During the first step, all the simulated values of the model are recorded after an artificial shock is provoked for the first two frames. During the second step, all the simulated values that occurred after the disturbance are compared with the simulated values that occurred before it (baseline). For example, in Figure 2.9, the simulated values of $RWRISTx(t)$ are depicted before the disturbance on the values of $RWRISTy(t-1)$ of 80%(in red) and after the disturbance on the values of $RWRISTy(t-1)$ of 80%(in blue). The disturbance on the simulated variables of $RWRISTx(t)$ is observed for 10 frames in total i.e. for 8 more frames and therefore 80% greater duration than the initial shock. A similar behaviour is also observed for $RWRISTy(t)$. The models adapt quickly after the application of the artificial shock, thus confirming their low sensitivity to external disturbances.

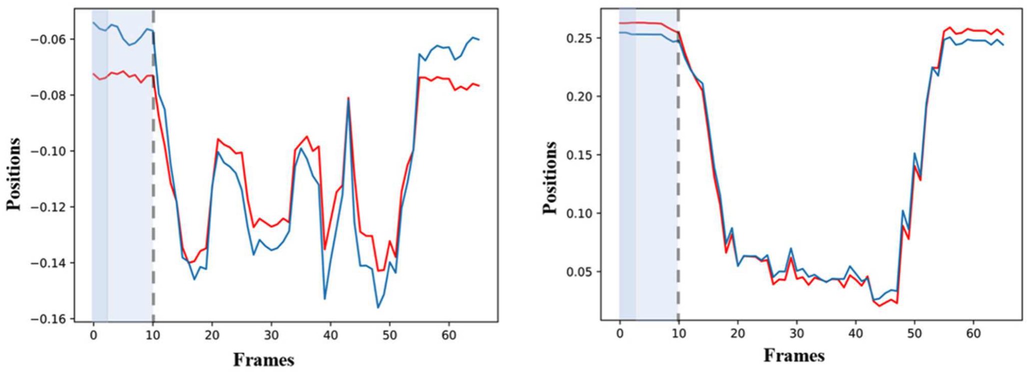


Figure 2.9: Left: Diagram of the simulated forecasted values of $RWRISTx(t)$ before the disturbance (red), and simulated forecasted values of $RWRISTx(t)$ after the shock (blue), for two frames, on the values of $RWRISTy(t-1)$ for 80% . Right: Diagram of the simulated forecasted values of $RWRISTy(t)$ before the disturbance (red) and simulated forecasted values of $RWRISTy(t)$ after the shock, for two frames, on the values of $RWRISTx(t-1)$ for 80% .

2.12 Summary of contributions

In this chapter, a generic methodology for professional gesture recognition is proposed, that uses multivariate time series as input. It is used to test cross-application scientific questions which are evaluated through various industrial scenarios.

In summary, the major scientific contribution of this chapter is the proposition of the GOM which represents human motion as the dynamic relationship of body entities (biomechanics) and their evolution in time (stochastics). Its statistical transcription into a simultaneous equation system facilitates multifarious possibilities for analysing body kinematic dexterity, increasing recognition accuracy and for forecasting motion trajectories.

Chapter 3

Human-Robot Collaboration

Contents

3.1 The Big Picture

A collaborative robot is an autonomous machine that is able to share workspace with the worker(s), without physical barriers, while still adhering to prescribed health and safety standards. Collaborative robotics has created the appropriate conditions for designing an HRC that can link human intelligence with the power of the robot by following one simple criterion: the complementarity of skills.

Following the principle of developing an external AI-based perception layer for collaborative robotics, the main contribution to HRC is the creation of a multi-user professional gesture recognition methodology. The recognition output is continuously communicated to the robot in order for it to adapt its behaviour according to the professional gestures and rhythm of the operator(s).

Here, two use-cases are considered from real-life situations, which were provided by the automotive assembly lines of the PSA Group: 1. that of door-assembly, which follows a co-presence scenario, and 2. that of motor-hose assembly, which follows a collaborative scenario.

From an algorithmic perspective, IMUs were used in the process of door-assembly in order to extract motion descriptors that were then used to train a hybrid HMM – DTW recognition engine. In the motor-hose assembly scenario, the body of the operator(s) is captured by a top-mounted depth sensor, then motion descriptors are extracted by applying geodesic distances on depth imaging and, in turn, motion time series are generated and provided as input to discrete HMMs for recognition of motion patterns (Figure 3.1). The performance of the AI-based perception layer was evaluated based on data gathered from multiple operators and by using additional information on gestures and actions that was provided by smart tools that were used by the operators. The gesture recognition algorithm was found to perform well also when a new operator joined the assembly line, by providing to the models only a few examples of his/her gestures.

The scientific contribution presented in this chapter was conducted by supervising the PhD of Eva Coupeté on 'Gestures and Actions Recognition for Human-Robot Collaboration in Assembly Lines' [Coupeté 2016], and received funding from the PSA Group of companies. A number of figures of this chapter are extracted from her thesis.

3.2 State-of-the-art

Industry 3.0 has rapidly transformed the assembly or production lines in business and industrial enterprises within only a few decades, thus enabling the switch from manual work to full automation. The manual work of the operators allowed them to make decisions and develop dexterous skills, thus also providing flexibility on production and processing lines. Nevertheless, standard industrial robots offered the possibility of massive automation, enabling work to be done in bulk with efficiency,

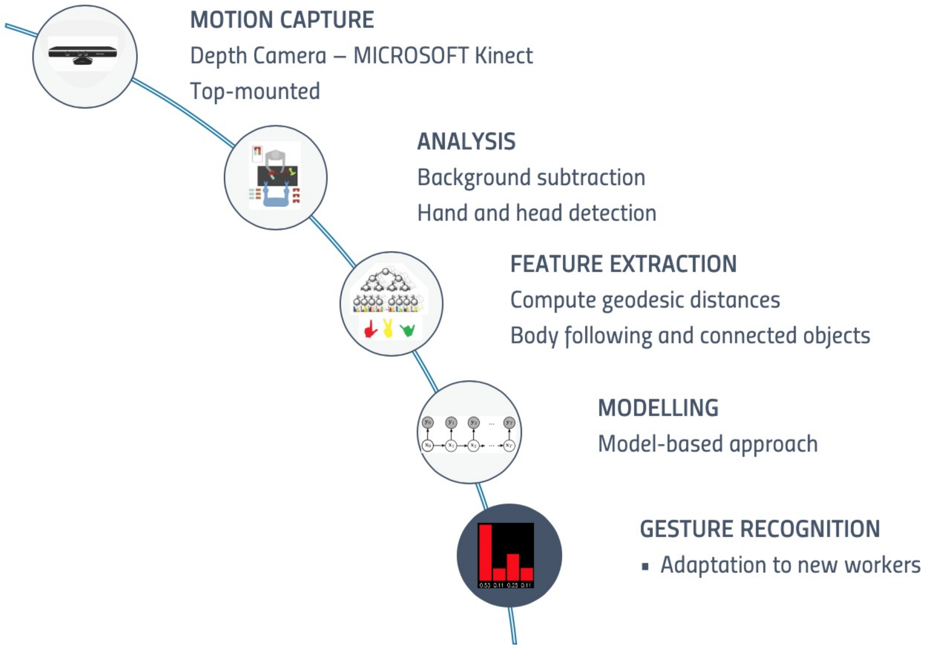


Figure 3.1: Methodology for professional gesture recognition in collaboration with a robot

speed and precision. However, this rapid conversion also introduced important constraints. Of these, the lack of flexibility to change the line is one of the most important i.e. the re-programming of the robot from scratch is very costly.

This industrial evolution also brought about an in-depth evolution in the use of the body of the operator(s). This evolution started from the human's professional gesture as a job initiator, passed through to the abundance of the traditional toolbox that machines could offer, until finally arriving at the complete replacement of the operators by the robots. Therefore, the human body as a tool, on the one hand, and the non-intelligent industrial robots, on the other hand, where there is no human in the loop, constitute two extreme situations (Figure 3.2).

With Industry 4.0, decision-makers realised that the deployment of robots that can work 'with' and not 'instead of' the humans would be beneficial for both the industry and the operators. Therefore, collaborative robotics began to open up new pathways for sharing the same space with the humans, while the robots could also be easily trained through only a few examples, thus generating important financial and economic gains. It is strange that despite these possible benefits and despite the very strong contributions from the scientific community and examples of deployments in industry, which consider the collaborative robot as a tool, only a few prototypes that consider the collaborative robot as a full partner with the operator have been reported.

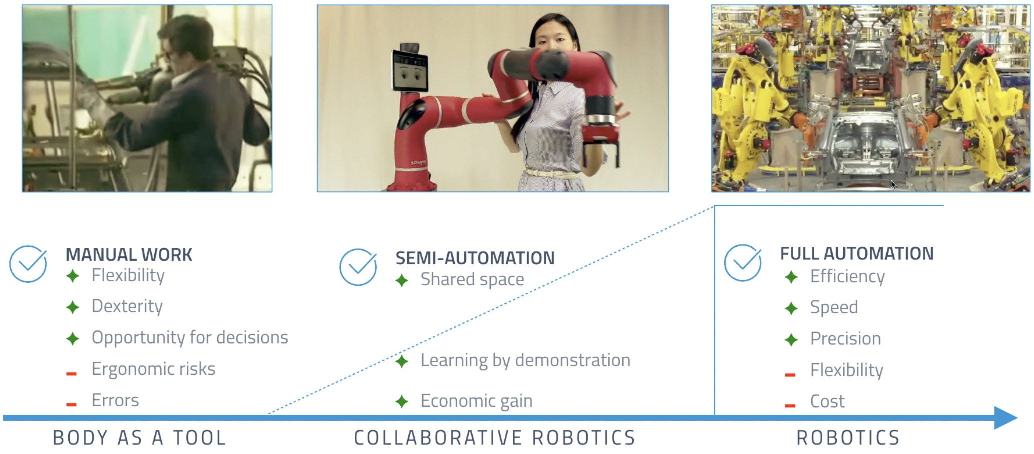


Figure 3.2: From manual work (body as a tool) to full automation then to semi-automation (collaborative robot as a tool)

3.2.1 Co-presence, cooperation and collaboration

Humans have the flexibility and the intelligence to be able to consider different approaches to solving problems, while robots can be more precise and consistent in performing repetitive and ergonomically risky tasks. Nowadays, mixed environments can be created which combine the cognitive skills of humans (intelligence, flexibility etc.) with the advantages of robots (high precision, repeatability etc.). In [El Zaatari *et al.* 2019] and [Kopp *et al.* 2020], the interaction between a human and a collaborative robot is distinguished according to the *physical working space*, whether separated or shared, to the *working time*, whether sequential or simultaneous, and to the *goal*, whether common or different, for the robot and the human operator respectively. *Physical contact* can also be considered as a criterion to characterise the interaction between the two in industrial robotics [Hentout *et al.* 2019a].

In manufacturing, various types of applications that involve robots, whether standard or collaborative, can be deployed. Standard non-intelligent industrial robots are put at the lowest level, where human and robot do not share their workspace and do not have any common task. The robot stops its task when a human presence is detected inside the cage.

Three types of interaction between a human and a collaborative robot are generally deployed in industry: *co-presence*, *cooperation* and *collaboration* [Schmidtler *et al.* 2015] (Figure 3.3). In a *co-presence* situation, human and collaborative robot share a part of the workspace, but do not work at the same task and physical contact is authorised only when the robot is stopped. The human operator adapts his/her rhythm and movements to the robot, whose velocity and trajectory are pre-defined.

In recent years, human and collaborative robots sequentially *cooperate* (also called sequential HRC), by fully sharing their workspace and working on the same task. Very often, the operator activates the task of the collaborative robot by pressing a button (Figure 3.4). Despite the sharing of the workspace, the human operator must adapt his/her behavior to the pre-defined temporal and spatial profile of the robot. From an industrial point of view, this cooperation between human and collaborative robots can be considered as the current baseline.

Finally, in a typical HRC situation, the collaborative robot should be responsive. It can be distinguished in physical Human-Robot Collaboration (pHRC) and touchless Human-Robot Collaboration (tHRC). In pHRC, there are operations which were intended to be without contact and instinctively the operator touches the robot, as well as operations where the operator touches the robot on purpose and the robot reacts in a particular way, depending on the direction and the force of the contact. In the first case, the robot reduces its velocity or stops its motion to avoid a collision [Michalos *et al.* 2015]. In the second case, the robot can either be used as a tool which extends the capabilities of the human operator (strength, preciseness etc.) or can be taught by demonstration in order to automatise a certain task. In tHRC, human and robot collaborate as colleagues or, alternatively, the robot assists the operator in his/her tasks, by combining HAI and motion sensors. Human action recognition is used to achieve contactless communication between the robot and the human operator. Therefore, the professional gestures are recognized by the robot, which adapts its behaviour accordingly. Nevertheless, this configuration is still very rare in the industry, mainly because only a few laboratory prototypes have been proposed.

3.2.2 Movement-based implicit and explicit adaptive interaction

Humans can interact with a collaborative robot through gestures either in an explicit way (e.g. to give a command) or implicitly (e.g. through an action that represents

	Shared time and workspace	**Common goal**	**Coordination**
Co-presence	✓	-	-
Cooperation	✓	✓	-
Collaboration	✓	✓	✓

Figure 3.3: Three types of interaction between a collaborative robot and an operator

a message for the robot) [Gildert *et al.* 2018, Hentout *et al.* 2019b]. Both types of interaction can generate a spatio-temporal adaptation of the robot according to the human's behaviour.

Explicit interaction for temporal adaptation can be achieved by pressing a button [Michalos *et al.* 2018] or by using a smartwatch as in the smart factory of BMW [1]. In this way, the operator informs the robot that a given task has been executed and this type of interaction can be categorised as a cooperation scenario. Force feedback and pointing gestures can also be used as a means for interaction [Cherubini *et al.* 2016].

Collision avoidance is ultimately the most frequent case scenario of spatial adaptation found today [Mohammed *et al.* 2017, Safeea *et al.* 2019]. Various use-cases of adaptation are also presented in the domain of assisting robotics. Therefore, robots can readjust their trajectories online and handle unpredictable incidents [Canal *et al.* 2018]. The robot adapts its behaviour both spatially and temporally, according to the anthropometrics and the rhythm of a human operator.

3.2.3 Professional gesture recognition for Human-Robot Collaboration

Gesture and action recognition can be used to improve the perception layer of the robot when collaborating with humans. Both Machine Learning (ML) [Liu & Hao 2019, Sharkawy *et al.* 2020a, Sharkawy *et al.* 2020b] and DL [El Dine *et al.* 2018, Heo *et al.* 2019] approaches have been implemented for collision avoidance or continuous gesture recognition [Tao & Liu 2013].

[1] https://roboticsandautomationnews.com/2017/03/04/bmw-shows-off-its-smart-factory-technologies-at-its-plants-worldwide/11696/

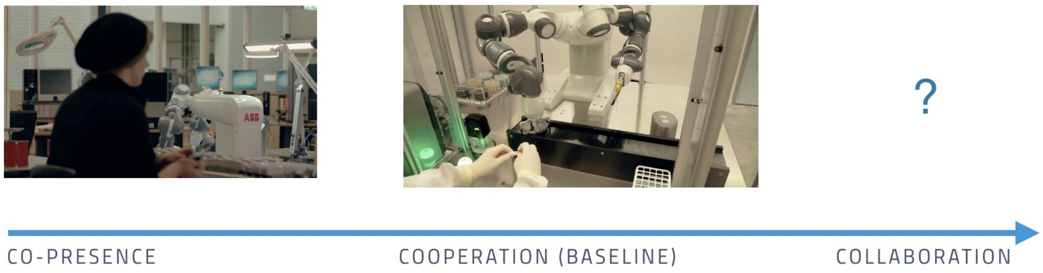


Figure 3.4: Most common examples of collaborative workspaces in industry. Co-presence: safe parallel work on different workbenches. Cooperation: sequential operation using an illuminated button. Green for robot's turn and orange for human's turn. By pressing the button we switch the turn. Collaboration: Only a few prototypes. Not yet exploited on a large scale in industry.

Reinforcement learning is used to minimize the risks of an incident in HRC. It encodes all task and safety requirements of the scenario into the settings of reinforcement learning, also taking into account components such as the behaviour of the human operator [El-Shamouty *et al.* 2020]. Interactive reinforcement learning is also used for programming the complete collaborative assembly process with the aim of reducing the effort needed by an expert engineer [Akkaladevi *et al.* 2018].

CNNs are also implemented for a multimodal HRC [Liu & Hao 2019], using LeapMotion for hand motion capturing, together with voice recognition. Human action recognition and tactile perception are used to distinguish intentional and incidental interactions when an incident involving physical contact occurs [Mohammadi Amin *et al.* 2020]. In this study, a 3D-CNN is used for action recognition, and a 1D-CNN for tactile detection with a Panda robot.

Discrete time Markov Chains are proposed to compute the probability that an incident may occur [Asaula *et al.* 2010], while the possible trajectories that a robot can follow, starting from a given position, can be predicted. Moreover, partially observed Markov decision processes are used to compute trust in the robot's decision-making [Chen *et al.* 2020]. Human trust is considered as a latent variable in a dynamic system that describes an effective HRC.

Most of the methods presented above are focused on specific factors in managing the shared workspace, without considering all the potential of the human as a partner of the robot. Safety and accident prevention are the most common goals in extracting information from the human.

3.3 Objectives beyond SoA

Leveraging on the above SoA, three objectives have been defined:

O1: *Development of a professional gesture recognition methodology*

Development of a methodology for online and continuous recognition of the situated professional gestures of the operator in automotive assembly lines.

O2: *Contactless collaboration between human and robot*

Development of an external perception layer for the robot that enables contactless collaboration.

O3: *Adaptability and dynamic cycle in collaborative cells*

Adaptability of the system when new operators join the cell and dynamic temporal collaboration depending on the rhythm of the operator.

The methodology developed for O1 contributes to the answering of Q1 (posed in Chapter 1, Section 1.2), by selecting the appropriate sensors, or the Internet of Things, and kinematic motion parameters that can be used for the accurate recognition of situated expert gestures, given the strong industrial constraints of the automotive assembly lines.

O2 aims to answer Q2 (posed in Chapter 1, Section 1.2) by exploring the possibility of replacing the current instrumental methods for collaboration (i.e. discrete commands by pressing a button) with a ‘gesture-following’ one.

O3 contributes to the answering of both Q1 and Q2 (posed in Chapter 1, Section 1.2). With regards to Q1, it further studies whether the same kinematic motion parameters can be used for adapting the machine learning models when a new operator joins the assembly line. As far as Q2 is concerned, the main contribution concerns testing whether the system of intelligence can be adapted to the rhythm of the operator and can distinguish the real professional gestures from unexpected movements e.g. looking at their phone etc.

3.4 Pose estimation

In the motor-hose assembly use-case, the depth sensor was mounted at the top of the scene to capture the actions of the operator, thus avoiding self or scene occultations. The main goal was to segment the scene, estimate the pose of the arms and localize his/her hands in real-time (Figure 3.5). The suggested approach is an extension of the work proposed in [Schwarz *et al.* 2012].

The first step of the scene segmentation is the exportation of body pixels from the depth image and the head localization. Since the operator is working in a limited workspace by using mainly his/her hands, we concluded that only the upper-part of the body would be sufficient for efficiently detecting his/her gestures. To achieve this goal, the head was detected by assuming that it was the closest to the camera object or body part. Following a number of anatomic considerations, the centre of the top of the head was estimated using the 10% of the highest body pixels, starting from the pixel closest to the camera (Figure 3.5). Moreover, pixels that were below the workbench of the operator were removed and the shoulders were localized on the depth map. Torso orientation is then calculated as the horizontal angle of the line connecting the shoulders.

It was therefore assumed that the hands are the farthest body parts from the head in the image. The geodesic distances were calculated between the top of the head and all the points of the upper body. A weighted graph was created using all the body pixels as nodes, by connecting each pixel with its neighbouring pixels. A weight was attributed to each pair of pixels, that is equal to the absolute difference of values between them, which can be considered to be approximately their difference in height, given the top-mounted depth camera. In order to avoid non-anatomic 'jumps' on the geodesic distances e.g. from the arm to the torso, connections with a very high depth difference were automatically dropped.

3.5 Feature extraction

As far as the motor-hose assembly use-case is concerned, the depth-processing algorithm provides the following features as output:

– 3D location of the top of the head;

– 3D locations for both hands;

– 3D geodesically shortest paths between head and hands.

Five different feature sets were tested, as presented in Figure 3.6, which

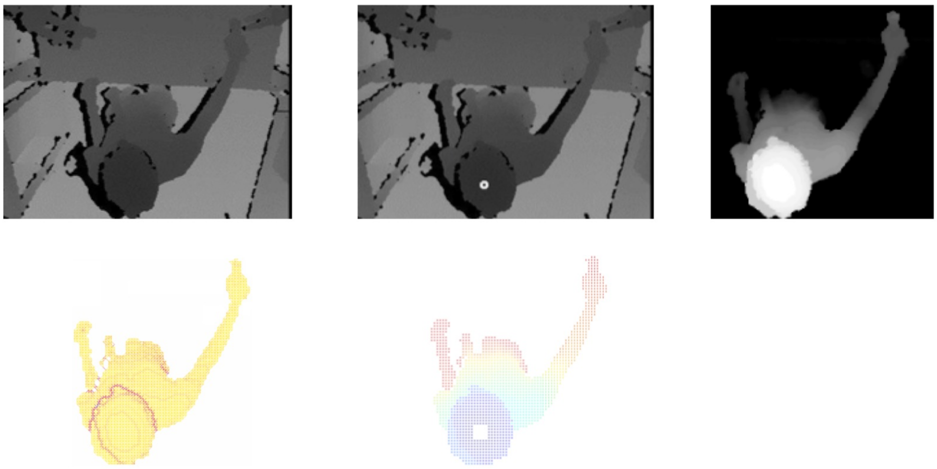


Figure 3.5: Basic steps for hand localization and upper-body pose detection. (1) depth-map from the top-view camera; (2) head localization; (3) torso exportation; (4) 2D graph of distances; (5) visualization of geodesic distance for each pixel of the upper-body

contain combinations between 3D locations for both hands and relevant upper-body postural information [Coupeté *et al.* 2019].

With regard to the *door-assembly* use-case, the IMU Animazoo IGS 120+ suit was used for the capturing of the human motions (Figure 3.7). This provided us with the angle of the segments of the body that followed a hierarchical model, where each angle was computed using as a reference point the angle of the previous body part. Thus, all of them used the sensor on the hips as the root.

3.6 Datasets and machine learning on time series

As was explained above, two different use-cases were studied: 1. *door-assembly*, which follows a co-presence scenario; and 2. *motor-hose assembly*, which follows a collaborative scenario. As a result, two datasets were created, one for each use-case, which contained the gestures that are presented in Figure 3.8.

In the *door-assembly* use-case, the duration of the gestures varied between

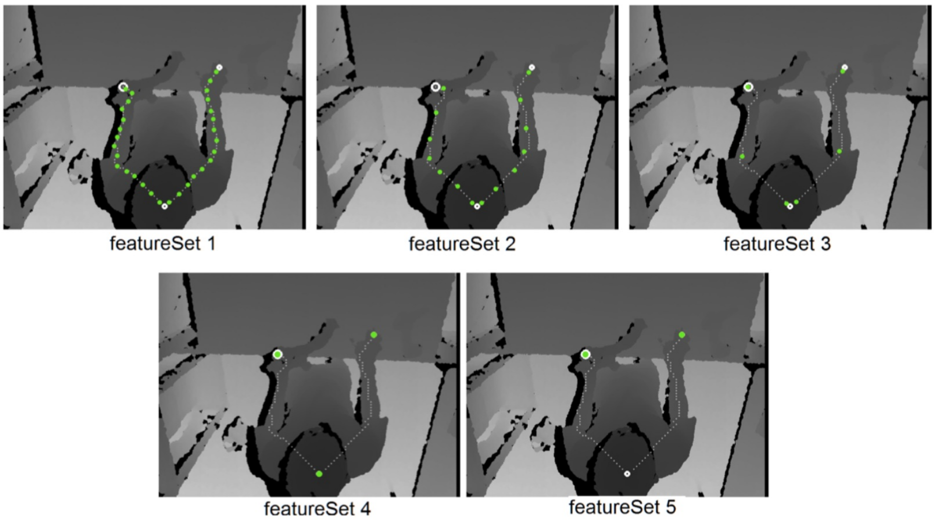


Figure 3.6: (1) 15 samples of each shortest path + 3D head and hands locations, (2) 7 samples of each shortest path + 3D head and hands locations, (3) 3 samples of each shortest path + 3D head and hands locations, (4) 3D head and hands locations and (5) 3D hands locations

8 - 10 seconds. Each worker performed 5 repetitions for each gesture, while each dataset contained observations of all the co-articulated gestures, without interruption. In the *motor-hose assembly* use-case, 2 female and 11 male 'naïve' operators were recorded, each with an average age of 47 years. Each of them executed between 20 - 25 assembly repetitions with 7 - 8 and 8 successive gestures respectively for each motion.

3.7 Template-based learning

It is a fact that the speed of the movement, together with other motion descriptors, can vary significantly from one repetition to another. For learning and recognition, we therefore used a DTW-based technique, known through its implementation in the 'Gesture Follower' tool [Bevilacqua *et al.* 2009]. This template-based method allows for one-shot learning while the online time-alignment between the template and the input gesture provided us with good recognition accuracy, even when there were only some minor variations in the performance of the same user. This approach was

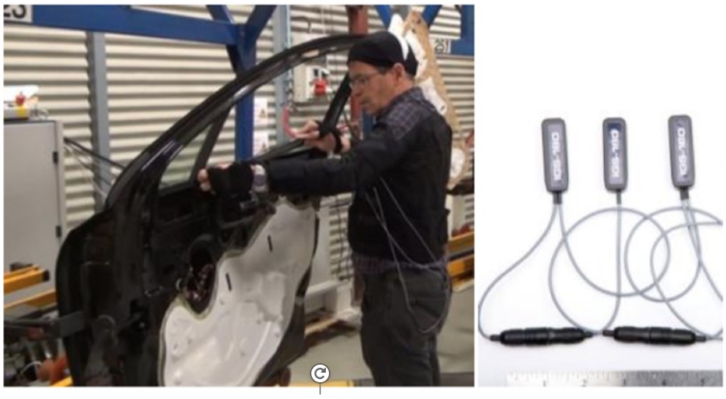


Figure 3.7: Operator wearing the Animazoo IGS120+ suit and dimension of the sensor

therefore deemed most suitable for implementation in the *door-assembly* use-case.

3.8 Model-based learning

Discrete HMMs were used to model the professional gestures of the *motor-hose assembly* due to these tasks being mainly action-driven, irregardless of the way the operator moves to execute them e.g. to obtain a specific position posture for taking a part from the robot. The continuous 3D locations provided by the geodesic distances between the head and the hands are then quantized to obtain discrete observations, through K-Means clustering. It involves partitioning observations into a fixed number K of clusters, where each observation belongs to the cluster with the nearest centroid, as described in [Coupeté *et al.* 2015, Coupeté *et al.* 2016]. In practice, each cluster corresponds to an approximate top-viewed posture. A gesture is therefore represented as a temporal sequence of cluster IDs, and thus as a sequence of approximate postures.

The cluster IDs are then used to train one discrete HMM per gesture class. Every feature vector that is extracted from a depth-image is quantized through the K-Means and the labels obtained are given to the discrete HMMs. A gesture is recognized when its associated HMM gives the highest probability for having generated the observations. The HMMs are trained with the Baum-Welch algorithm while the Forward algorithm is used for recognition. They are both implemented in

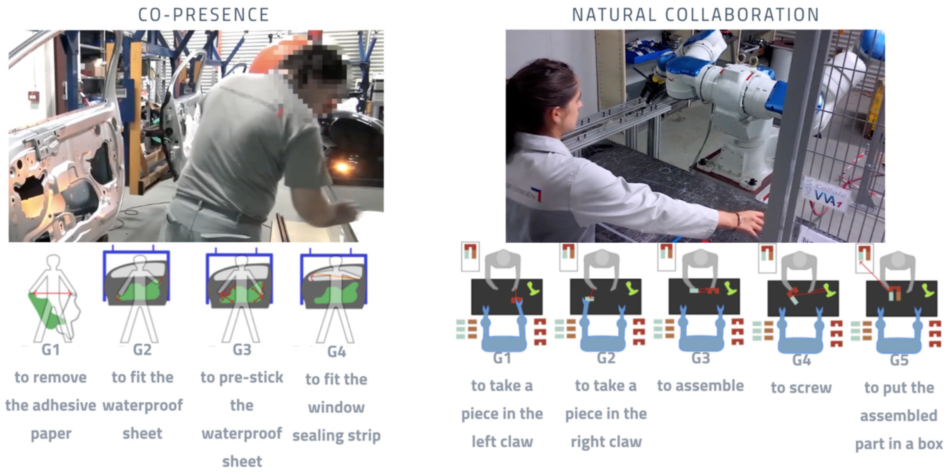

CO-PRESENCE
G1
to remove the adhesive paper
G2
to fit the waterproof sheet
G3
to pre-stick the waterproof sheet
G4
to fit the window sealing strip
NATURAL COLLABORATION
G1
to take a piece in the left claw
G2
to take a piece in the right claw
G3
to assemble
G4
to screw
G5
to put the assembled part in a box

Figure 3.8: Gesture vocabularies for door-assembly (left) and motor-hose assembly (right)

the GRT[2] open library. Figure 3.9 illustrates our methodology.

3.8.1 Instrumenting the tools

Instrumenting the tools of the worker can be beneficial for obtaining a greater recognition accuracy through leveraging the data generated from their online and connected behaviours. From the surroundings of the human body, additional information is obtained by placing inertial sensors on the screwing-gun of the operator. Then, a 'late-fusion' of this data is used and acts as an elimination criterion for gestures where the algorithm 'hesitates' e.g. between 'screwing' and 'assembling' (Figure 3.9). The screwing-gun is therefore only supposed to move when the worker is using it to screw together two parts of the motor hose.

In practice, by instrumenting the screwing-gun, two types of conflict can be resolved: a. when the algorithm outputs a screwing gesture while the screwing-gun is not moving, if the likelihood of the HMM for 'screwing' is above a threshold then this gesture is recognized, otherwise the output is 0; and b. when the algorithm outputs a non-screwing gesture while the screwing-gun is moving, if the likelihood of the HMM for 'screwing' is above a threshold then 'screwing' is recognized instead

[2] https://github.com/nickgillian/grt

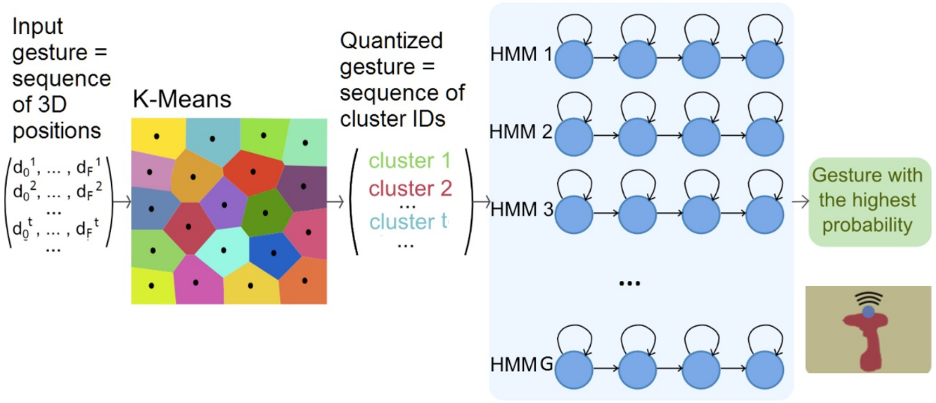

Input gesture = sequence of 3D positions
K-Means
Quantized gesture = sequence of cluster IDs
cluster 1
cluster 2
...
cluster t
...
HMM 1
HMM 2
HMM 3
...
HMM G
Gesture with the highest probability

Figure 3.9: Gesture recognition pipeline: input gesture (left) is a temporal sequence of feature vectors of the same dimension F; each continuous-valued feature vector is quantized by K-means into a discrete-valued 'approximate posture' label (middle); the temporal sequence of successive posture labels obtained are fed one after the other into G discrete HMM (1 per gesture class); for each time-step, our system outputs the most probable current gesture class, by selecting the HMM which has current maximum likelihood.

of 'non-screwing', otherwise a 'non-screwing' gesture is recognized.

3.9 Classification and recognition assessment

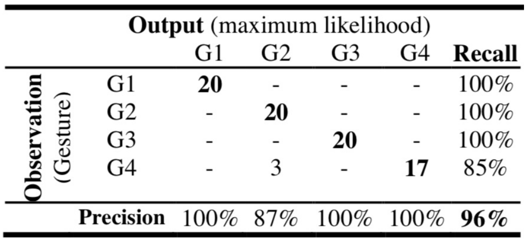
The tests that are used for the evaluation of the performance of the algorithm are a. the *jackknife* and b. the *80%-20%*. Jackknifing consists of systematically recomputing the statistical estimate, leaving out all the data of one operator, for testing, and using the data of all the remaining operators for training (one operator per iteration), in this way doing all the possible combinations. Therefore, in each iteration, one operator is considered as 'unknown' by the algorithm. The 80%-20% test, on the other hand, consists of randomly separating the whole dataset into two separate datasets, where the one for training contains 80% of the initial dataset, while the one for testing contains the remaining 20%. Thus, it estimates the performance of the algorithm, while it is trained on at least some data from all the operators.

3.9.1 Classification of isolated motion patterns

3.9.1.1 Door-assembly use-case

To estimate the accuracy of the DTW-based approach to classifying isolated motion patterns, the 'jackknife method' was used. In practice, this means that one of the five datasets is left out and used for training for each iteration, until all the datasets have been used once, while the remaining datasets are used for testing.

If we analyse the results according to the Figure 3.10, G3 ('pre-sticking the waterproofing-sheet') is recognized perfectly, while the accuracy for G4 ('to fit the window sealing strip') is the lowest, with only 85% of recall. The excellent results of G3 can be explained by the fact that the operator makes circular movements without walking at all, in contrast with the other 3 gestures where s/he uses the whole workspace. Moreover, G3 has very low intra-class variability. With regard to precision, G2 ('to fit the waterproofing-sheet on door') has the lowest performance, with only 87%. As far as G2 and G4 are concerned, they both have similar postures that generated confusion in their HMMs.



		Output (maximum likelihood)				
		G1	G2	G3	G4	**Recall**
Observation (Gesture)	G1	**20**	-	-	-	100%
	G2	-	**20**	-	-	100%
	G3	-	-	**20**	-	100%
	G4	-	3	-	**17**	85%
	Precision	100%	87%	100%	100%	**96%**

Figure 3.10: Precision and Recall for isolated motion patterns (*use-case: door-assembly; sensor: IMUs; method: DTW)*

3.9.1.2 Motor-hose assembly use-case

The performance of each feature set was evaluated using jackknifing, as shown in Figure 3.11. The best recognition was given by the feature set of 3D hand locations.

The results can be explained through a simple observation: the more we add information that is not related to body parts that contribute to the effective gestures, the more the recognition rate decreases.

15 samples head location hands locations	7 samples head location hands locations	3 samples head location hands locations	head location hands locations	hands location
65%	**70%**	**72%**	**74%**	**79%**

Figure 3.11: Gesture recognition rate per set of features (*use-case: motor-hose assembly; sensor: depth; method: HMMs - K-Means*)

Then, the parameters of the HMMs and K-Means were optimised by using the two hands 3D locations as feature sets. Different combination parameters, mainly K for the number of clusters of K-Means and S for the hidden states of the HMMs, were tested by using the jackknife as a criterion for measuring their performance. According to the Figure 3.12, the more K increases, the more the description of the human motion is detailed and the better the accuracy is, until K gets values between 20 and 25. For K=25 and S=12 the recognition rate is optimised.

		Number S of HMM states					
		5	**7**	**10**	**12**	**15**	**20**
Number K of K-means clusters	**10**	74%	75%	73%	72%	74%	73%
	15	76%	78%	78%	79%	78%	79%
	20	77%	80%	77%	78%	79%	78%
	25	76%	77%	79%	**82%**	81%	80%
	30	77%	78%	78%	80%	80%	79%

Figure 3.12: Clusters for K-Means vs states for HMMs (*use-case: motor-hose assembly; sensor: IMUs*)

3.9.2 Early and continuous real-time recognition of gestures

3.9.2.1 Door-assembly use-case

For the *door-assembly* use-case, the DTW method was used, as described in Section 3.7. According to Figure 3.13, there is almost always a template that generates a maximum probability that is usually near 1. Nevertheless, there are fluctuations in the transition phase between gestures due to a co-articulation phenomenon (blue curve at the bottom of Figure 3.13). Thus, there are cases where for a short time period there are false maximum probabilities from a non-relevant template, that could justify the assumption that the algorithm might need more data as input before being 'aligned' with the relevant gesture. More precisely, the fluctuations are longer for G4 mainly because the beginning of G4 and G2 is quite similar, while only in the end does G4 differentiate from G2. To address the issue of false recognition, a sliding window, with a length equal to 1/3 of the duration of the shortest gesture (1.64 seconds), was implemented.

Within the window, only probabilities above 0.7 are taken into consideration and the algorithm returns the template that gave the most maximum probabilities for at least 75% of the timestamps. Otherwise, the algorithm returns '0', meaning that no gesture is recognized with enough confidence. In this way, most cases of false recognition are eliminated. According to Figure 3.13, we see that the algorithm returns '0' at the beginning of both G2 and G3, as well as in the middle of G3. Nevertheless, the error with G4 is not eliminated. Finally, the average delay for a gesture to be recognized by the algorithm is 3.4 seconds.

3.9.2.2 Motor-hose assembly use-case

For the motor-hose assembly use-case, hybrid HMMs - K-means were used, as is described in Section 3.8. In early and continuous gesture recognition, accuracy depends also on the length of the temporal sliding window and whether or not the tools of the operator are instrumentalised with motion sensors.

However, *which lengths for the temporal sliding window optimise $\overline{P}$ and $\overline{R}$ and how much are they improved when instrumenting the screwing-gun?* Several comparisons of various lengths (from 0.5 to 2 seconds), were made using both the jackknife and the 80%-20% tests. In the jackknife test, for a temporal window with a length of 1 second while the screwing-gun is instrumentalised, the metrics are optimised for $\overline{P}$=84% and $\overline{R}$=77%, compared to $\overline{P}$=76% and $\overline{R}$=74% without any instrumentalisation of the screwing-gun. Thus, by instrumenting the tool, the values of the metrics are improved by +8% for $\overline{P}$ and +3% for $\overline{R}$, which, in conclusion, means that the algorithm becomes much more precise and accurate. In the 80%-20%

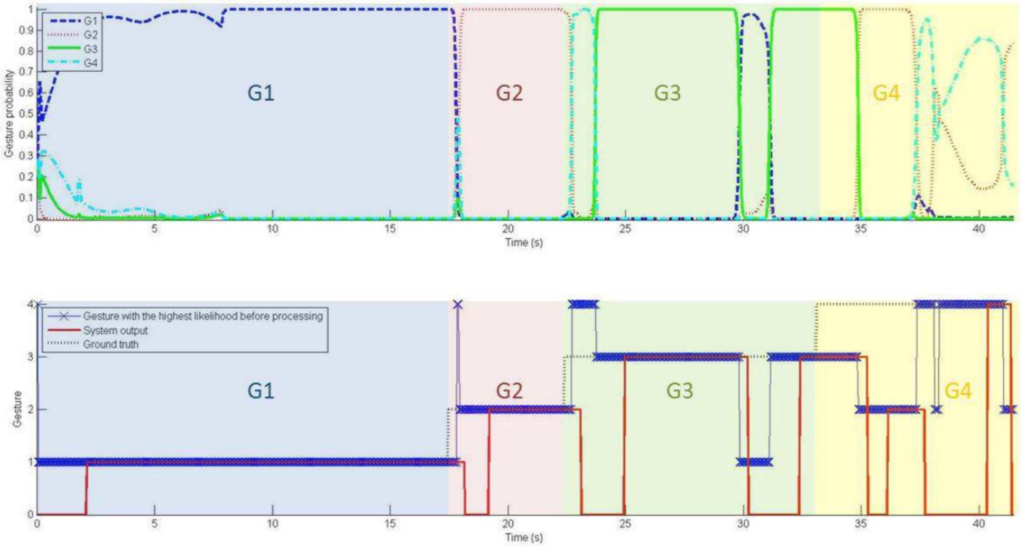


Figure 3.13: Top: Output of probabilities in real-time without sliding window. Bottom: Maximum probabilities only (blue), recognition after the sliding window (red), and the ground truth (black). G1 (in blue background), G2 (in pink background), G3 (in green background) and G4 (in yellow background) are sequentially executed (*use-case: door-assembly; sensor: IMUs; method: DTW*)

test, when instrumenting the screwing-gun, $\overline{P}$=82% is optimised for a length of 1.5 seconds, while $\overline{R}$=85% for both lengths of 1 and 1.5 seconds. Thus, medium-sized temporal sliding windows optimize $\overline{P}$ and $\overline{R}$ metrics, while the $\overline{R}$ is better using the 80%-20% test mainly because of the fact that the operator is ‘ ’known’ to the algorithm.

According to Figure 3.14, we see the algorithm is able to recognize the gesture before its end. The delay in recognition is between 1 and 1.5 seconds for the duration of gestures that vary between 1.5 and 3 seconds. For G3 and G4 specifically, recognition occurs at 0.9 and 0.6 seconds respectively, before the end of each gesture.

In industry, having a new operator join the assembly line is a routine task; however, the accurate recognition of his/her professional gestures, without applying major modifications to the dataset, can prove to be a research challenge. A partial answer to the question of how to address this challenge is given through the tests

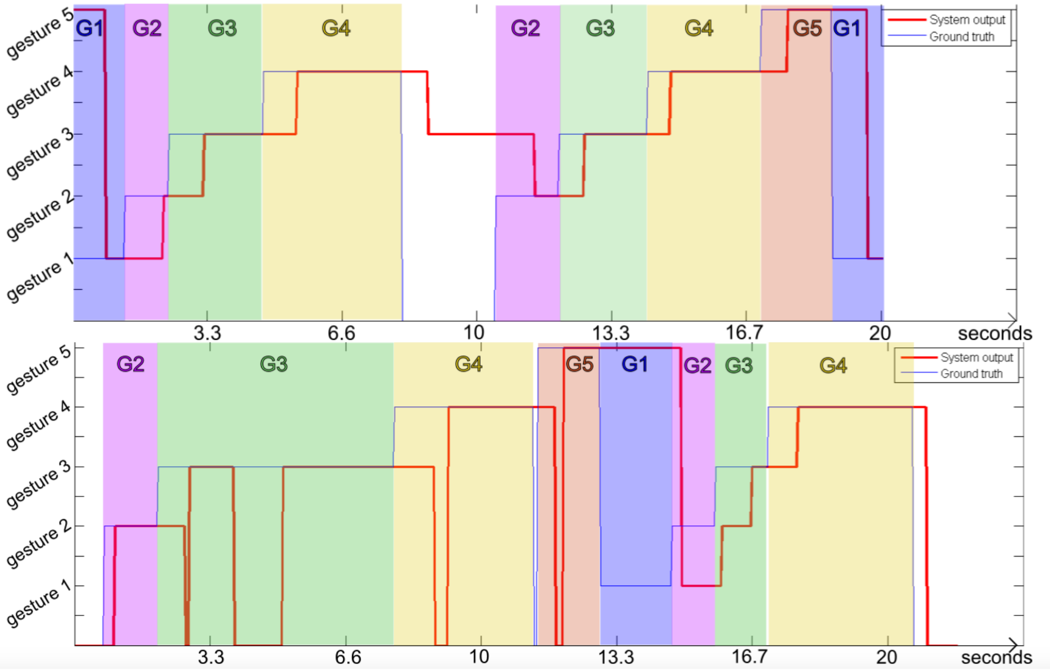


Figure 3.14: Two examples of early and continuous gesture recognition. Top: Maximum probabilities in real-time. Bottom: Maximum probabilities and '0' outputs when probabilities are below 0.7. Blue line and colors on the background: ground truth; Red line: algorithm output. (*use-case: motor-hose assembly; sensor: depth; method: HMMs - K-Means*)

that were performed to assess the accuracy of the algorithm: Jackknife vs 80%-20%. That is because this comparison taught us that when the algorithm 'knows' the operator, the accuracy becomes higher.

Nevertheless, there remains the question of *how much motion data from the new operator would maximise this accuracy?* To provide an answer to this question, a set of gestures were progressively added while $\overline{P}$, $\overline{R}$ and F were computed for every set added, with a temporal sliding window equal to 1 second. Among all the experiments executed, the metrics were maximised for 15 datasets ($\overline{P}$=89%, $\overline{R}$=89% and F=89%), where a +5% is obtained for all $\overline{P}$, $\overline{R}$ and F compared to their values when only 1 set of gestures is added, as well as +5% for $\overline{P}$ and +12% for $\overline{R}$ when absolutely no data from the worker are used for training ($\overline{P}$=84% and $\overline{R}$=77% from the jackknife test previously mentioned), or +13% for $\overline{P}$ and +15% for $\overline{R}$ without instrumenting the screwing-gun.

However, *would a significant improvement be possible when fewer gesture sets are added?* In factories, the deployment of an algorithm that involves a large number of human recordings is usually limited. According to the top-right plot in Figure 3.15, a small addition of gesture sets e.g. less than 5, can improve $\overline{P}$ and $\overline{R}$ by up to +9% and +3% respectively. When more than 5 sets of gestures are added, the impact on accuracy converges at approximately 1%, which is a significantly lower value than when there are up to 5 sets (top-right of Figure 3.15).

In order to measure the performance of the algorithm on mono-user datasets as well, a number of experiments were conducted, while an increasing number of examples were added to the training dataset, together with a constant testing dataset. It was found that the more the number of additional gesture sets grew, the better the recognition performance was, such as in the case of 1 to 7 sets, where r increases from 80% to 89%, thus by +9%. However, the improvement becomes much slower when more than 7 sets are added. According to the plots at the bottom of Figure 3.15, $\overline{R}$ values from mono-user datasets are generally better than those of the multi-user datasets, while they become similar for 15 newly added sets of gestures. Nevertheless, $\overline{P}$ values from both multi- and mono-user datasets are globally close, while above 10 sets, $\overline{P}$ values from the multi-user adapted training are better than those with the mono-user dataset. This would seem to indicate that the training of the algorithm using a multi-user adapted training has the potential to be more robust than a mono-user training.

3.10 Summary of contributions

In this chapter, a tHRC methodology was developed and implemented for two industrial use-cases, that of door-assembly (co-presence) and that of motor-hose assembly (collaboration) scenarios.

In summary, three major scientific contributions are presented. The first contribution is the development of a vision-based methodology for recognising the situated professional gestures of operators. The algorithm receives depth imaging as input and continuously outputs the actions of the operators online. The second scientific contribution is the 'on-the-fly' temporal adaptation of the robot's behaviour, according to the operator's behaviour, which opens a broad pathway for considering the robot as a partner instead of a tool. Finally, it has been proved through this study that building external HAI-based perception layers onto the robot, gradually increases its understanding of the actions of its partner, thus contributing towards a more natural collaboration.

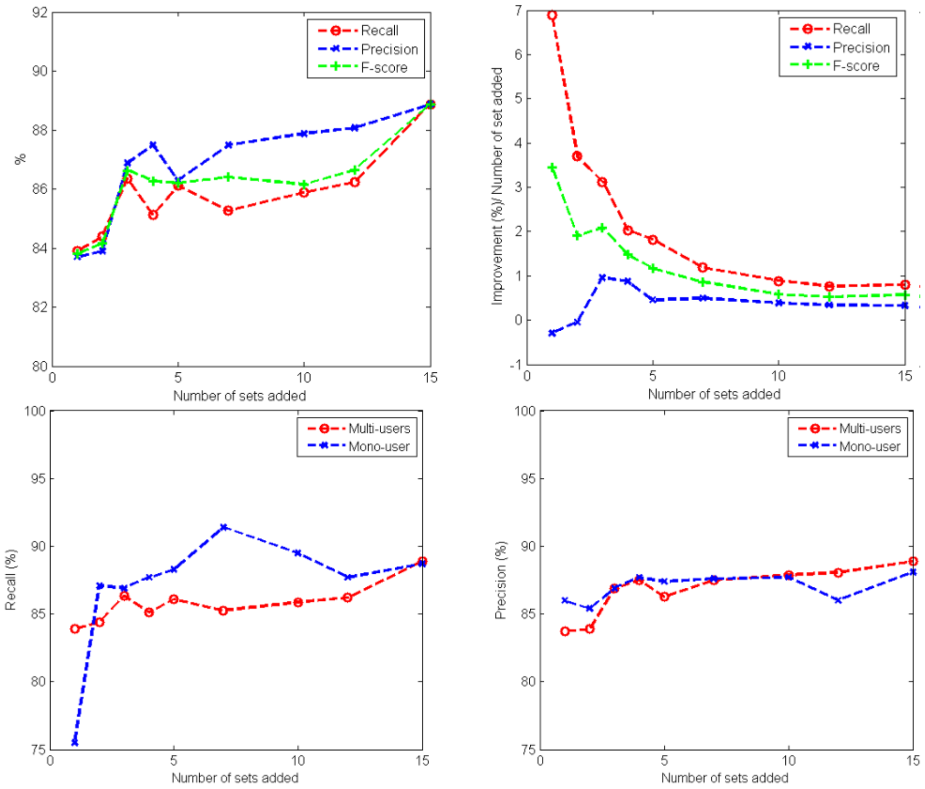


Figure 3.15: Training adaptation when a new operator joins the assembly line. Top: $\overline{P}$, $\overline{R}$ and F values (left) and their improvement (%) (right) as a function of progressively added sets of gestures for a multi-user training. Bottom: Comparison of $\overline{P}$ (right) and $\overline{R}$ (left) values when additional gestures sets are put in the multi-user (with red lines) and mono-user (with blue lines) training datasets. (*use-case: motor-hose assembly; sensor: depth; method: HMMs - K-Means*)

Figure 3.16 illustrates my scientific contributions to tHRC. The more human-centric the information that is extracted, the more the perception of the robot is enriched. The more its perception is enriched, the more it can anticipate the human's behaviour and accordingly adapt its own, thus rendering the collaboration more natural. In other words, in order for the robot to become a partner of the human, it needs to be able to understand not only its task but also its colleague. The introduction of collaborative robots into industry (level 0) is the baseline (fast easy training but no adaptation during collaboration) and still the most frequent case scenario found today. However, the button that activates the process of the robot can be replaced by the detection of the position of the operator's hand when s/he is ready to receive a part from the robot (level 1). Both spatial and temporal profiles remain constant and predefined. Moreover, human action and gesture recognition enable the temporal adaptation of the robot according to the rhythm of the operator (level 2). I am currently working on the development of dynamic spatio-temporal profiles that receive the positions and actions of the operator as input parameters, as well as on the extraction of movement analytics that will permit the monitoring of the operator (level 3).

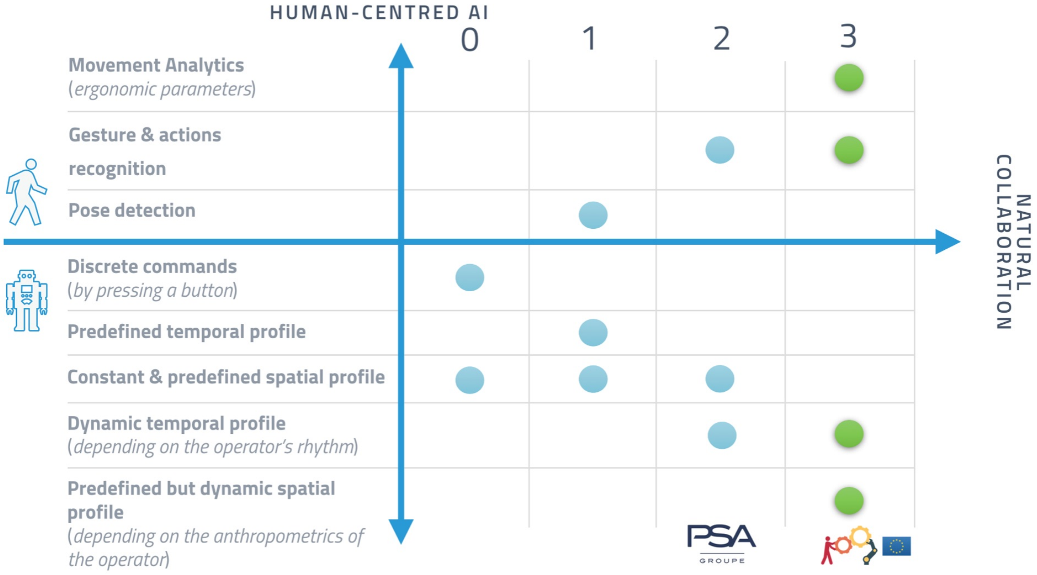

HUMAN-CENTRED AI
0
1
2
3
NATURAL COLLABORATION
Movement Analytics
(ergonomic parameters)
Gesture & actions recognition
Pose detection
Discrete commands
(by pressing a button)
Predefined temporal profile
Constant & predefined spatial profile
Dynamic temporal profile
(depending on the operator's rhythm)
Predefined but dynamic spatial profile
(depending on the anthropometrics of the operator)
PSA GROUPE

Figure 3.16: Overview of scientific contributions (in blue) for a natural Human-Robot Collaboration and ongoing work (in green). '0' stands for the current baseline in the industry.

Chapter 4

Computer-mediated sensori-motor human learning

Contents

4.1 The Big Picture

This chapter describes my scientific contribution to the in-person transmission of expert movement skills to apprentices. An HAI-driven methodology was developed, in which the computer acts as mediator in sensori-motor human learning. The methodology proposes a performance mode for capturing and modelling the expert know-how, as well as a learning mode for its transmission to the apprentice (Figure 4.1).

In the performance mode, the motions of the expert are captured using vision-based sensors. Motion descriptors are extracted from the signal and stochastic learning is used to model kinematic elements of his/her expertise and to recognize them on a unknown signal. In learning mode, the apprentice imitates the expert gestures and motion descriptors are extracted from his/her signal and compared with the expert descriptors in real-time. When the apprentice deviates from the expert gesture, both sonic and optical sensori-motor feedback is generated by the computer in order for him/her to adjust the gesture according to that of the expert. The deployment of the methodology is articulated around three scientific hypotheses concerning the contribution of HAI and motion capturing in: 1. extracting expert knowledge from the recordings; 2. the self-training of the apprentice after having received an in-person training session from the expert and 3. generating sensori-motor feedback to assist with human learning.

The performance of the approach is evaluated in the use-case of the wheel-throwing art of pottery. For this purpose, two expert craftsmen were recorded: one from Vallauris (considered to be France's clay capital) and Mr Thodoris Galigalidis from the Ceramic School of the Therapeutic Centre KETHA ITHAKI in Thessaloniki, Greece.

The scientific contributions presented in this chapter were conducted by supervising the PhD of Alina Glushkova on 'Gesture recognition technologies in managing movement skills. Sensori-motor feedback as a gamification mechanism' [Glushkova 2016], funded by the i-Treasures FP7 project. A number of figures of this chapter are extracted from her thesis.

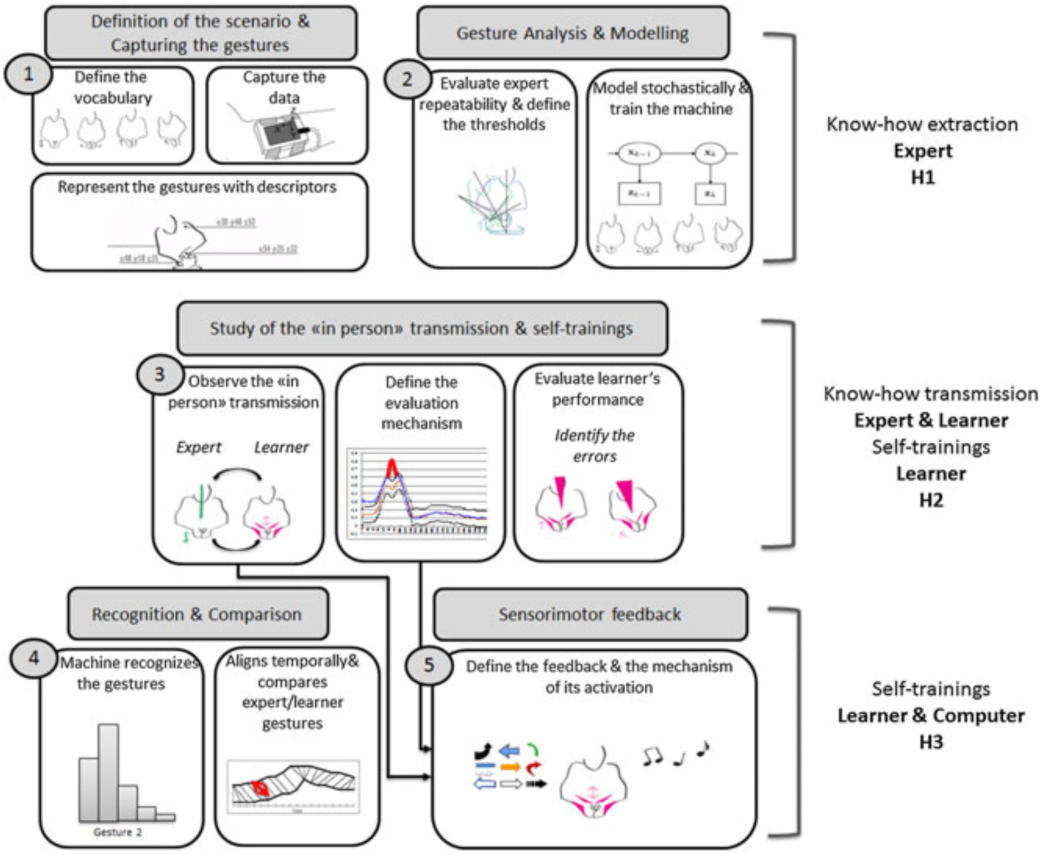

Definition of the scenario & Capturing the gestures
1
Define the vocabulary
Capture the data
Represent the gestures with descriptors
Gesture Analysis & Modelling
2
Evaluate expert repeatability & define the thresholds
Model stochastically & train the machine
Know-how extraction
Expert
H1
Study of the «in person» transmission & self-trainings
3
Observe the «in person» transmission
Expert
Learner
Define the evaluation mechanism
Evaluate learner's performance
Identify the errors
Know-how transmission
Expert & Learner
Self-trainings
Learner
H2
Recognition & Comparison
4
Machine recognizes the gestures
Gesture 2
Aligns temporally & compares expert/learner gestures
Sensorimotor feedback
5
Define the feedback & the mechanism of its activation
Self-trainings
Learner & Computer
H3

Figure 4.1: Methodology for capturing, modeling and recognition of expert gestures with the aim of computer-mediated human learning

4.2 State-of-the-Art

Movement-based interactive systems for sensori-motor learning generate augmented feedback which supports human learning. These systems are based on a feedback mechanism that collects motion data as input and computes the deviation/variance between the reference gesture (in this case, the expert gesture) and the incoming gesture (in this case, the apprentice gesture). Such systems are being used in various application domains, such as in rehabilitation [Kitago & Krakauer 2013] or in vocational training, but particularly in contexts where the transmission of skills occurs in person [Dimitropoulos *et al.* 2018].

4.2.1 Sensori-motor human learning principles

The learning of motor skills was studied quite thoroughly and extensively by psychologists, physiologists and ethnologists in the 20th century and continues to be so. Newell divides the motor skills into perceptual, cognitive, communicative and other categories [Newell 1991]. Humans develop a sensori-motor intelligence that allows for motor control, coordination, and action.

According to Piaget, embodied intelligence is acquired through the senses, through experiences and consequent ambient reactions [Piaget 1976], where a mapping between motor and sensory parameters is created [Wolpert *et al.* 2011]. For example, the contextual cueing phenomenon occurs during the learning process where visual features of objects, tasks, or the environment are associated with kinematics [Makovski 2018]. Failure in learning enriches sensori-motor mapping, according to Roger Schank, a cognitive psychologist and AI-theorist [Schank 1997].

Sensori-motor feedback aims to support the process in which the human learns to anticipate his/her environment and to eliminate kinematic errors. It is based on a priori modelling of human movements [Manitsaris *et al.* 2014a]. *Motor skill acquisition* constitutes a fundamental phase in the learning process where a new motion pattern is transmitted to the apprentice. It is a dynamic process where the apprentice learns both how to select single movement elements and combine them fast and accurately [Diedrichsen & Kornysheva 2015]. Shmuelof defines motor skill acquisition as an improved trade-off between speed and accuracy [Shmuelof *et al.* 2012]. Another fundamental phase in the human learning process is *motor adaptation*. Wolpert defines it as the process by which the motor system adapts to perturbations in the environment [Wolpert *et al.* 2011].

The body mechanisms that are activated in the two sensori-motor learning phases vary significantly. Changes of a high amplitude occur in the way human movement is executed during the skill acquisition phase, while micro kinematic and kinetic adjustments occur in the adaptation phase.

4.2.2 Machine learning supporting sensori-motor human learning

Various machine and deep learning techniques can be applied in sensori-motor human learning. Recently, Caramiaux published a short review on adaptation capabilities of machine learning for human movement modelling [Caramiaux *et al.* 2020], focusing on parameters adaptation on probabilistic models, and transfer or meta-learning or adaptation through reinforcement learning. Nevertheless, template-based methods have also proved their scientific value with evidence, when the learning occurs in-person (e.g. expert/apprentice), providing an online characterisation of the temporal evolution of the incoming gesture by rescaling the reference gesture.

4.2.2.1 Probabilistic models

Nowadays, stochastic and probabilistic approaches, such as HMMs, Gaussian Mixture Model (GMM)s or SS, are considered as 'classic methods' for gesture recognition or generation, as previously described in Section 2.2.2.

GMMs are widely used in robotics for teaching the robot to adapt its movement parameters when new target parameters are set [Calinon 2016]; they are also used in gesture control of sound by teaching the parameters of a new user with a single example. HMMs are applied when it is important to take into account the internal states of a gesture, whether for recognition [Françoise & Bevilacqua 2018] or generation of stylistic walking movements [Tilmanne 2013]. SS is used to model the spatiotemporal dynamics of the human body by updating the state parameters of the GOM and 'specifying' the HMM-based recognition using the confidence bounds of an expert performance [Manitsaris *et al.* 2020] (presented in Chapter 2). It is also used to model movement variations, such as speed, scale and rotation, as presented in [Caramiaux *et al.* 2014].

Therefore, parametric adaptation of probabilistic models allows for micro-kinematic adaptation to tolerated movement variations, when restricting them, using either a single example or a few examples.

4.2.2.2 Transfer and meta-learning with Deep Neural Networks

In transfer learning, an initially trained neural network that learns movement features (or embeddings) on a source domain is adapted to a new or similar knowledge domain, thus providing it with good features.

Temporal convolutions are used for transfer learning, whether for improving the classification accuracy when both the source and target domains consist of motion data from the same users, in different periods [Rad & Furlanello 2016], or for automatically synthesizing character movements from given trajectories, by mapping low level human motion to high level parameters that are easily configurable by the new user [Holden *et al.* 2016]. Spatio-temporal convolutions are also used for transfer learning, whether for gesture recognition, when only a few examples per class are available [Kikui *et al.* 2018], or for improving classification accuracy [Côté-Allard *et al.* 2019]. An RNN is trained offline on the human motion transitions and then online using recursive least square errors for adapting the robot behaviour to that of the human gestures [Cheng *et al.* 2019].

According to Caramiaux, there are still research challenges concerning how the size of the training datasets affects transfer learning and whether a catastrophic 'forgetting' could occur after successive transfers [Caramiaux *et al.* 2020].

In meta-learning, or learning-to-learn, the goal is to improve the learning algorithm given the experience obtained through various learning episodes [Hospedales *et al.* 2020].

Meta-learning is applied in teaching a collaborative robot to perform a set of actions using one-shot imitation learning with a regressor against the output actions, resulting in an equal performance between seen and unseen demonstrations [Duan *et al.* 2017]. Moreover, a model-agnostic meta-learning approach that is compatible with any model trained with gradient descent is introduced by Finn [Finn *et al.* 2017] and extended by Yu to cover one-shot imitation learning by a robotic arm [Yu *et al.* 2018]. Finally, similar approaches are also applied to human motion forecasting for unseen tasks when big annotated datasets are available [Gui *et al.* 2018].

4.2.2.3 Deep reinforcement learning

Deep reinforcement learning assumes that machines can learn from their actions, similar to the way humans learn from experience, using trial-and-error interactions with their environment.

Many strategies for training a reinforcement learning algorithm are proposed. One strategy is to initialise the reinforcement learning network using imitation learning [Kober *et al.* 2013]. Another strategy consists of inversing the reinforcement learning by deriving cost functions from demonstrations [Finn *et al.* 2016].

Furthermore, learning a movement from expert performances constitutes a major challenge for the in-person transmission of motor skills. Generative Adversarial Imitation Learning (GAIL) can go some way towards overcoming the problems by discriminating between expert motion trajectories and artificially generated ones [Ho & Ermon 2016] and completes the process through reinforcement learning where the acceptable trajectories are defined by previous successful performances [Guo *et al.* 2018]. When GAIL and reinforcement learning are simultaneously combined, the algorithm learns both faster and more accurately [Zhu *et al.* 2018].

Nevertheless, the training of a reinforcement learning architecture is often very slow and may generate unnatural motion trajectories [Caramiaux *et al.* 2020].

4.2.2.4 Similarity measurement using temporal rescaling

Although a pair of time series might appear to produce shape and amplitude similarity (e.g. expert and learner motion data), they can be considered to be out-of-phase, and, in such cases, classical distance measurement fails.

By temporally warping the motion time series, similarity measurement functions can be applied to characterize the distance between the sequences by contemplating temporal elastic shifting [Folgado *et al.* 2018]. DTW and Longest Common Subsequence (LCSS) are applied to compensate for non-linear distortions and measure the similarity between the time-series. Nevertheless, there are a number of challenges related to this approach such as in cases where an improper x-axis warping occurs (phenomena of 'singularities') trying to express amplitude variability in the y-axis [Keogh & Pazzani 2001]. A number of variants of the DTW have been developed to address this issue, such as Derivative Dynamic Time Warping (DDTW) or Weighted Dynamic Time Warping (WDTW), but they still depend strongly on the nature of the data (e.g. whether or not the signal has a higher degree of information on its first derivative). The implementation of the DTW over a sliding window, as with the approach of [Gillian *et al.* 2011] or [Folgado *et al.* 2018], can overcome the alignment challenges of the DTW by reflecting a feature-to-feature or window-to-window similarity. A distance metric, such as the Euclidean or the Mahalanobis, can then be applied to the two motion time-series, over 3D motion data, or as a standard deviation over the axis.

Time-warping related approaches are usually based on one-shot learning since they consider the first data sequence as reference and the second as incoming gesture. Although they suffer from a quadratic time and space complexity, they have proven their applicability in the human motion analysis domain, such as in music and dance [Gillian *et al.* 2011, Ferguson *et al.* 2014], rehabilitation [Jégo *et al.* 2013] and sensori-motor learning of professional gestures [Glushkova & Manitsaris 2015].

4.2.3 Augmented feedback in situation-based sensori-motor human learning

Humans make use of various types of information during motor control [Adams 1971, Rigal 2003], which is linked to their proprioception. In situation-based learning, the augmented or extrinsic feedback is information that is provided to the learner, through the use of an external source, whether that is via the expert/teacher or is computer-mediated [Astill & Utley 2008, Schmidt & Wrisberg 2008]. The types of augmented feedback are analytically described in Figure 4.2

According to [Sigrist *et al.* 2013], concurrent augmented feedback usually helps the novice to quickly understand the structure of the gesture, without having any cognitive overload. It also helps experts/teachers in the sense that teaching specific details of the movement can be a complex task, especially when the expert also has to monitor and provide the learner with feedback concerning automated but incorrect movement. Furthermore, the more the performance of the learner improves, the more the frequency of the concurrent feedback should be reduced, or simply switched to terminal feedback. According to the multiple resource theory of Wickens [Wickens 2008], audiovisual feedback enhances the perception level of the learner. The distribution of the information to the 2 modalities is appropriate since vision is usually highly loaded (e.g. when manipulating tools or objects). Moreover, when the goal of the task is to learn a specific trajectory, visuohaptic feedback is reported to be beneficial (e.g. in reducing spatial errors) [Ruffaldi *et al.* 2009].

Finally, general conclusions purporting that sonic feedback is systematically better than visual feedback are not consistent, especially if the sound is not properly designed [Sigrist *et al.* 2013]. Multimodal feedback seems to be beneficial for the learning of complex tasks. Nevertheless, because of the great technical effort required to organise scientific studies for a large number of real-life situations, there is a lack of systematic evaluation of the efficiency of the feedback, given the gesture categories [Wolpert *et al.* 2011].

4.3 Objectives beyond SoA

The purpose of this research was to study whether computer-mediated sensori-motor feedback can have a positive influence on the process of human learning motor skills. Thus, three research hypotheses were proposed (Figure 4.1):

H1 - Knowledge extraction: *A machine can learn to recognize the kinematic parameters of expert technical gestures*

To confirm or refute this hypothesis, a number of experiments were implemented to test whether a machine is able to recognize expert gestures with high accuracy after having had training on a number of gesture examples. The relationship established between the expert and the learner during the in-person transmission, as well as the difficulties the learner faces during self-training have been taken into consideration.

H2 - In-person transmission and self-training: *Machine learning and gesture recognition can contribute to the evaluation of the gestural performance of a learner in a self-training situation without providing any feedback*

To confirm or refute this hypothesis, a number of metrics that evaluate the learning progress are defined, these being based on both the spatial and temporal parameters

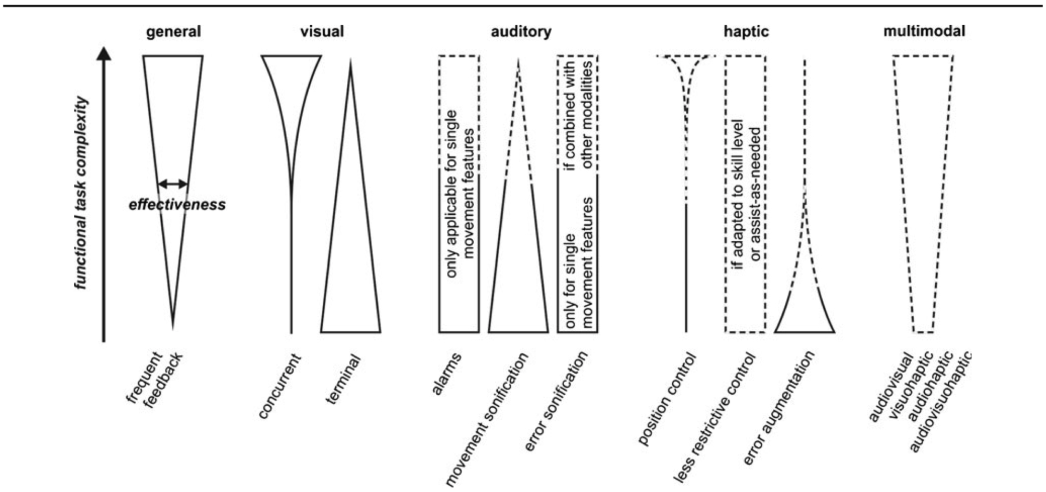


Figure 4.2: Feedback categorisation for experimentally confirmed (solid) or hypothesized (dashed) effectiveness to enhance motor learning depending on functional task complexity, according to [Sigrist *et al.* 2013], the broader the shape, the more effective the strategy is.

of the gestures. The experiments were implemented after in-person transmission and without the presence of the expert.

H3 - Impact of computer-mediated sensori-motor feedback on human learning *The computer-mediated sensori-motor feedback has a positive impact on human self-training.*

To confirm or refute this hypothesis, the metrics from H2 were used to evaluate whether the gesture performance of the learner improved when computer-mediated sensori-motor feedback was provided.

4.4 A methodology for training the computer to assist the human learning of movement skills

The identification of the embodied knowledge was based on the implementation of semi-conductive interviews with the experts in order to collaboratively define the scenario and its specifications as well as the vocabulary of gestures (Figure 4.1). The type of motion sensors that were compatible with the scenario were also defined and used for recording the expert performing a sufficient number of repetitions for all the gestures of the vocabulary. Then, the motion data, whether positions or rotations for a number of body joints or segments, were pre-processed and normalized for use as a dataset for training machine learning models.

In the second methodological layer, Human-centred Artificial Intelligence (HAI) methods and techniques were used to build a perception layer for the computer that would allow it to recognize the gestures of the vocabulary when an unknown example was given to it [Manitsaris *et al.* 2014b, Glushkova & Manitsaris 2015]. More precisely, machine learning models were trained on the motion data of experts. The ability of the computer to recognize unknown examples of the gesture vocabulary e.g. from an apprentice, is derived from the ability to successfully mathematise the kinematic parameters of the expert gestures (Hypothesis H1). In order for the computer to be able to continuously understand whether the learner is correctly performing the gesture or not, the *intra*-expert tolerance per gesture had to be calculated. The goal was to evaluate the degree of repeatability for the expert and to explicitly define the confidence bounds, beyond which the gesture e.g. of an apprentice, is considered as *bad*, according to the expert data (see Chapter 2). To do so, one option was to use the average or maximum of the standard deviation of the expert repetitions. In this research work, the one-shot-learning approach introduced by [Bevilacqua *et al.* 2009, Bevilacqua *et al.* 2007] was used to test the hypotheses. It consists of a hybrid approach between DTW and HMMs that is based on template-based learning and allows the use of a single gesture to define a gesture class. Nevertheless, an HMM formalism was used to compute, in real-time, measures between the template and the incoming data stream.

In the third methodological layer, there was a 3-step procedure where: A. the relationship between the expert and the learner during the *in-person transmission* was studied (expert and learner working together); B. the mechanism for evaluating the learner was defined and C. once the *in-person transmission* was completed, the self-trained performance of the learner was evaluated (learner alone), during which no assistance was provided to him/her. In order for the digital learning-by-doing experience to be as natural as possible, the observations from Step A were taken into consideration in designing the sensori-motor feedback mechanism [Piaget 1976]. The various steps in this methodological layer contributed to the testing of Hypothesis H2.

In the fourth and final methodological layer, the learner experienced real-time gamification in a self-training situation. The computer received as input the gesture of the learner and recognized it (using HMMs for the pilot pottery use-case). Then, it temporally aligned and compared it (using DTW for the pilot pottery use-case) with the reference gesture of the expert. During the alignment, the evaluation mechanism was activated to verify whether the kinematic parameters of the learner's gestural parameters were within the confidence bounds that were defined by the expert performance. When the performance of the learner was outside the confidence bounds, the sensori-motor feedback mechanism was activated by the computer, during which implicit or explicit sonic/optical indications were communicated to the learner in order for him/her to apply micro-movement adjustments and correct his/her gestures 'on-the-fly'. The self-trained performance of the learner (learner alone), while receiving online sensori-motor feedback, was evaluated again (Hypothesis H3).

4.5 Capturing the motion of expert craftsmen: the wheel-throwing pottery use-case

In order to study if expert gestures in wheel-throwing pottery could be captured, analysed, and modelled through motion sensors and HAI, we conducted an experiment with the initial participation of two potters, as described in a previously published paper [Manitsaris *et al.* 2014b]. The scenario selected was the creation of a simple bowl (18/23 cm diameter), which was completed through 4 phases, each of them containing 4 or 6 gestures, depending on the object size and the quantity of clay (Figure 4.3). The object of the first potter is smaller (18-20 cm in diameter, 10 cm in height, using approximately 1.3 kg of clay) and his 4 main gestures correspond to the 4 phases. The object of the second potter is bigger (20-23 cm in diameter, 13 cm in height, using 1.75 kg of clay). Thus, the second potter has more clay to manage and he is paying more attention to shape refining. Each gesture was executed 5 times and recorded with the use of IMUs.

IMUs provide data about the rotations of body segments, which are mainly Euler angles. From a cognitive point of view, it is difficult for humans to interpret angles as a form of feedback for controlling their gestures. For example, informing the learner that his/her hand should be rotated by 32 degrees cannot be easily assimilated from a pedagogical point of view. For this reason, a third expert potter was recorded on the same gesture vocabulary with a Microsoft KINECT sensor, which provides the joint positions in 3D space.

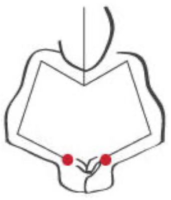
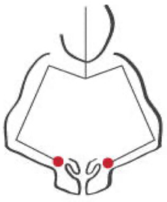
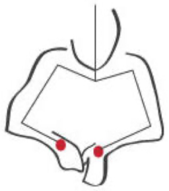
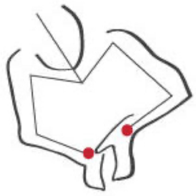
						
4 basic gestural phases	**P_1** **Centering and bottom opening**		**P_2** **The raise**	**P_3** **The first configuration**	**P_4** **The final configuration and removing**	
Potter A 4 gestures	G_1^A Centering and bottom opening		G_2^A The raise	G_3^A The first configuration	G_4^A The final configuration and removing	
Potter B 6 gestures	G_1^B Centering the clay	G_2^B Opening the bottom	G_3^B The raise	G_4^B The first configuration	G_5^B The final configu ration	G_6^B Removi ng the object

Figure 4.3: Main phases and gestures for the creation of a ceramic bowl

4.6 From in-person transmission to computer-mediated sensori-motor human learning

4.6.1 Analysing in-person transmission and self-training: the role of vision, hearing and touch senses

In the beginning, a mirroring system ('me-to-you' observation system) was established, with the expert having an active role and the learner having a passive role (*vision* is the main sense). The expert performs the gestures in real conditions, with clay. The learner interprets the visual information and virtually imitates the 4 gestures introduced by the expert, with the same rhythm and speed. Then the roles are flipped. The expert observes the learner performing the gestures and assists him/her with oral and sometimes visual instructions (*vision* and *hearing* are the main senses). Finally, the learner starts performing the gesture in real conditions. Examples of oral instructions are: "push the clay higher", "press the clay to centre it", "close your hands to get a smaller object diameter". The expert also uses physical contact to adjust the movement of the learner in the case of errors (*vision* and *hearing* and *touch* are the main senses). During the in-person transmission, the learner is continuously assisted and evaluated by the expert potter.

Once the in-person transmission is completed, the learner practises alone. In pottery, as well as in other manual arts and professions, the movement skills are acquired only through practice and experience. In this self-training step, 11 beginner-learners (average age 26.2 years old; 10 right-handed and 1 left-handed; 6 women and 5 men; with a high degree of familiarity with technological devices such as computers and smartphones, but no previous experience in using movement-based interactive systems) participated in the experiment. Their gestures were recorded using the Microsoft KINECT sensor.

4.6.2 Designing computer-mediated sensori-motor learning

The expert instructions can be categorized according to 2 criteria: A. their *function*, that is often linked to the moment they intervene and B. the *senses* that the learner should activate. Moreover, the instructions can be *explicit*, thus giving a clear message (e.g. "open your hands"), or *implicit*, that require interpretation by the learner.

The design of the computer-mediated feedback is based on the expert instruction typology (Figure 4.4) and the analysis of expert/learner relations during the in-person transmission. Particular focus is placed on the correction and evaluation instructions during the natural in-person transmission and their transfer into the computer-mediated sensori-motor learning. Both implicit and explicit optical computer-mediated feedback is designed and activated by the evaluation mechanism when there is a significant deviation in the learner's hand distance. The implicit visualization consists of curves/waves that vary depending on the deviation value, so that the greater the deviation is, the longer the curve. The goal of the learner is to have a thin deviation line, equating with zero. For example, when the wave concerning horizontal distance deviation appears on the right, it means that the learner's hand distance is greater than that tolerated and she or he must 'close' their hands to reduce the distance. Also, explicit optical instructions are provided, showing the general distance trajectories within which movements are to be performed (Figure 4.5).

Acoustic feedback is also provided when the deviation is not tolerated, using the sound of a bell to attract the learner's attention. Finally, at the end of the gesture, evaluation feedback is also given to the learner, to provide him/her with a global picture of his/her performance, in order to reinforce his/her motivation. This takes the form of a global score that is given to the learner based on his or her temporal success percentage (100% minus percentage of temporal failure).

anticipation	**correction**		**evaluation**
acoustic	optical	acoustic	acoustic
explicit	implicit	explicit	
		implicit	

Figure 4.4: Expert instruction typology

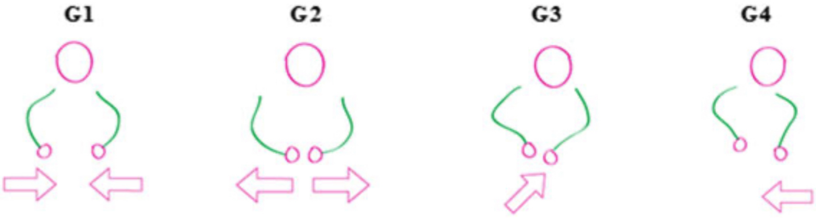


Figure 4.5: The four indications used as explicit optical feedback

4.7 Modelling kinematic parameters of expert gestures

To evaluate the performance of the algorithm, the metrics introduced in Section 2.9 were also used to test the hypothesis on knowledge extraction (H1). They were generated following the jackknife procedure. The performance of the algorithm was excellent for Potter 1, $\overline{P}$=100% and $\overline{R}$=100% (3D Euler angles for body segments), almost excellent for Potter 2, $\overline{P}$=96% and $\overline{R}$=97,5% (3D Euler angles for body segments) and very good for Potter 3 $\overline{P}$=93,5% and $\overline{R}$=93,7% 3D joint positions). Thus, H1 is confirmed since the algorithm was able to recognize kinematic parameters of expert skills when an unknown example was presented to it (i.e. jackknife).

Measuring the repeatability for every expert potter is an important step before defining how much the algorithm should tolerate the performance of the learner when imitating this specific expert. Thus, in order to identify which body parts contributed most to the execution of the effective gestures, a Principal Component Analysis (PCA) was applied, based on previous work presented in [Volioti *et al.* 2014]. In addition to the 3D positions of the 9 articulations of the body, the distance of the hands/wrists was also tested. According to the Principal Component Analysis (PCA), hand distances constituted the principal component in the dataset. Then, the standard deviation of the distance was computed after having temporally aligned all the repetitions of the same gestures using the DTW. Standard deviation was computed for each frame and coordinate axis and the maximum value for each axis was taken as a tolerance threshold (λ). For example, Potter 3 had a very high level of repeatability since while executing G2 twice, he might permit himself to have a difference of only 1.02 cm on the x-axis. The z-axis has the most important deviation from one repetition to the other, because the potter was moving his chair or changing his position in front of the camera. Thus, the z-axis was not used for the evaluation of the learning process.

4.8 A mechanism for evaluating the learner and activating the feedback

After the modelling of the expert kinematic parameters, the learner was invited to use the same sensors and algorithm as the expert in order to do self-training. Then, the algorithm presented in real-time with a recognition probability using the learner's gesture as input and the expert gestures as reference. The higher the probability, the better the performance of the learner is.

During the recognition process, both time series, from the expert and the learner, were temporally aligned. Then, spatio-temporal deviation measurement was applied. As far as the temporal deviation is concerned, its duration should be as close as possible to that of the expert. According to the expert, the learner should assimilate the speed and rhythm of each gesture and the acceptable temporal deviation is that of 5 seconds.

As far as the spatial deviation is concerned, the tolerance threshold λ is added to the expert Euclidean hand distance DE. Every time the learner's hand distance DL exceeded the greatest tolerated distance $DE + \lambda$, or was less than the least tolerated distance $DE - \lambda$, his/her performance went beyond the confidence bounds and was thus considered to have deviated from the expert model. Thus, the following controls were considered:

$$\begin{aligned}
&if\ DL_x < (DE_x - \lambda)\ then\ (DL_x - (DE_x - \lambda))\\
&if\ DL_y < (DE_y - \lambda)\ then\ (DL_y - (DE_y - \lambda))\\
&if\ DL_x < (DE_x + \lambda)\ then\ (DL_x - (DE_x + \lambda))\\
&if\ DL_y < (DE_y + \lambda)\ then\ (DL_y - (DE_y + \lambda))
\end{aligned} \tag{4.1}$$

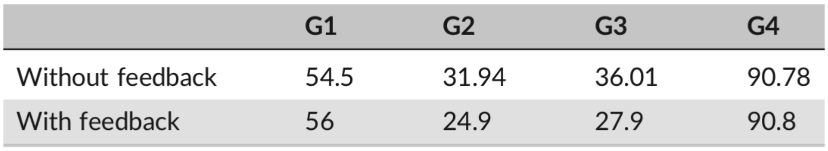
The total deviation of a gesture is the addition of instant deviations, thus giving a final score per axis.

4.8.1 Temporal deviation with and without computer assistance

To evaluate the temporal deviation and learning progress of the apprentice, the difference between the average expert $\overline{DurL_i}$ and learner $\overline{DurEx_i}$ duration per gesture i for all of the 11 learners was computed:

$$\sum_{i=0}^{11}(\overline{DurL_i} - \overline{DurEx_i}) \tag{4.2}$$

According to Figure 4.6, for G2 and G3, by indicating on the computer monitor the current and desirable duration of the gestures, the learners were able to better approach the temporal goal. The gesture with the biggest deviation was G4, mainly because it contained an important number of movements of great amplitude (taking a wire, removing the object from the wheel etc.) which required experience.

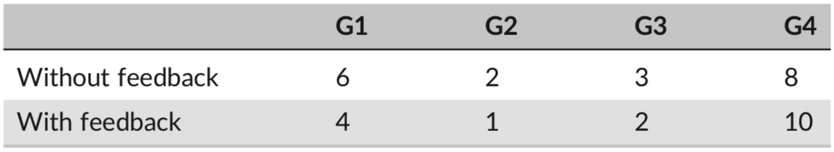


	G1	G2	G3	G4
Without feedback	54.5	31.94	36.01	90.78
With feedback	56	24.9	27.9	90.8

Figure 4.6: Sum of average temporal deviations with and without the time feedback

According to Figure 4.7, the number of learners who had important deviations from the experts' average was reduced for three of the four gestures. G4 was an exception because the learning of the articulation of the different movements inside G4 required additional time for practice.

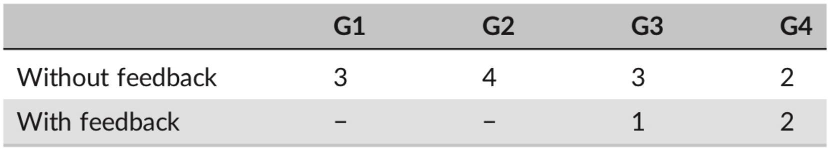
The expert stated that it is important for the learner to have temporal homogeneity while repeating the same gesture. However,the learner may perform the



	G1	G2	G3	G4
Without feedback	6	2	3	8
With feedback	4	1	2	10

Figure 4.7: Number of learners with $\overline{DurL_i} - \overline{DurEx_i} > 5$ sec with and without time feedback

same gesture with a very different duration (from 25 - 45 seconds, or even 50 seconds for G2 etc.). According to Figure 4.8, the use of sensori-motor feedback significantly reduced temporal variation among the gestural performance of the learners for three of the four gestures. In particular, for G1 and G2, all the learners had a temporal deviation per gesture of $<$5 sec. Thus, by providing sensori-motor feedback of the time counter, the learners reached a better perception of the desirable duration.

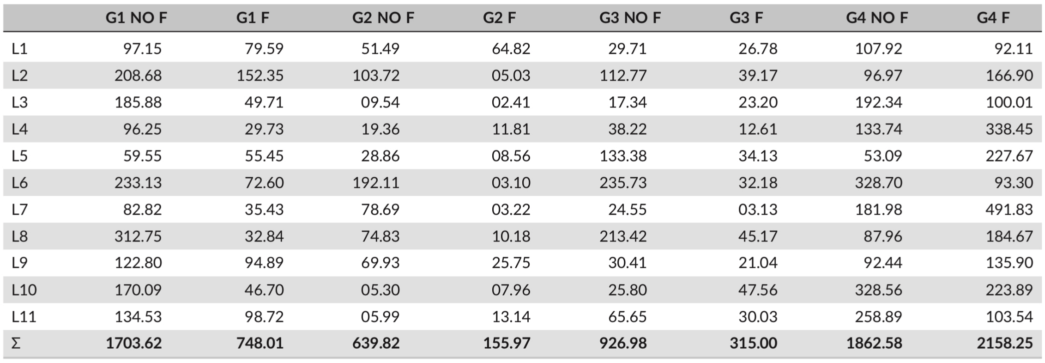


	G1	G2	G3	G4
Without feedback	3	4	3	2
With feedback	–	–	1	2

Figure 4.8: Number of learners with a temporal variability higher than 5 sec

4.8.2 Spatial deviation with and without computer assistance

By comparing the average deviations of the hand distances of the 11 pottery learners (with and without computer assistance) with the expert hand distances, we concluded that the learner's kinematic performance improved with the use of sensori-motor feedback (Figure 4.9). Three of the four total gestures improved, while the interpretation of visual and sonic feedback seemed to have helped students to understand the correct gesture trajectories.



	G1 NO F	G1 F	G2 NO F	G2 F	G3 NO F	G3 F	G4 NO F	G4 F
L1	97.15	79.59	51.49	64.82	29.71	26.78	107.92	92.11
L2	208.68	152.35	103.72	05.03	112.77	39.17	96.97	166.90
L3	185.88	49.71	09.54	02.41	17.34	23.20	192.34	100.01
L4	96.25	29.73	19.36	11.81	38.22	12.61	133.74	338.45
L5	59.55	55.45	28.86	08.56	133.38	34.13	53.09	227.67
L6	233.13	72.60	192.11	03.10	235.73	32.18	328.70	93.30
L7	82.82	35.43	78.69	03.22	24.55	03.13	181.98	491.83
L8	312.75	32.84	74.83	10.18	213.42	45.17	87.96	184.67
L9	122.80	94.89	69.93	25.75	30.41	21.04	92.44	135.90
L10	170.09	46.70	05.30	07.96	25.80	47.56	328.56	223.89
L11	134.53	98.72	05.99	13.14	65.65	30.03	258.89	103.54
Σ	**1703.62**	**748.01**	**639.82**	**155.97**	**926.98**	**315.00**	**1862.58**	**2158.25**

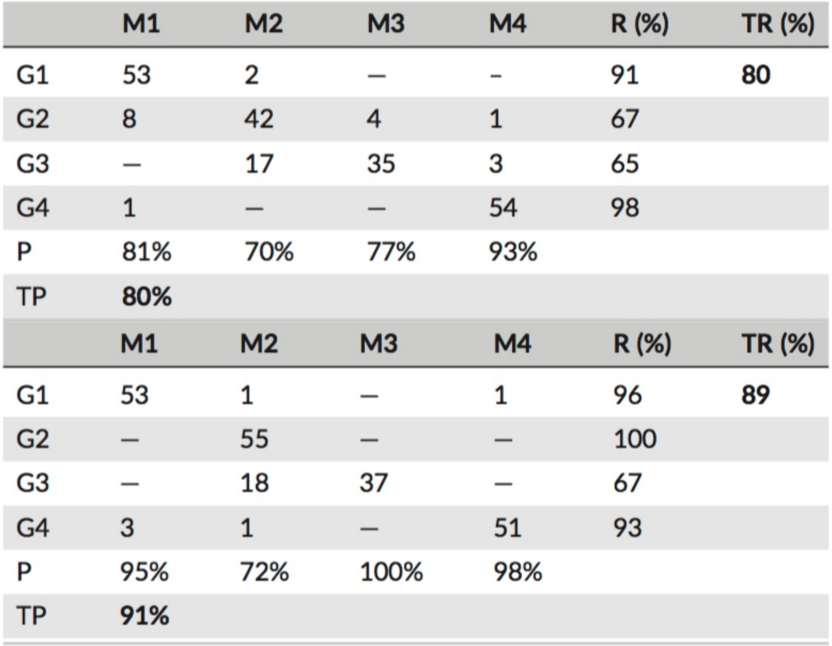
Figure 4.9: Average spatial deviation in centimetres on x and y axes compared to the expert, with and without feedback

4.8.3 Recognition accuracy as a criterion for evaluating the learning curve

Leveraging on the results presented so far, we can assume that any improvement in the recognition accuracy of the learner's gestures during the computer-mediated sensori-motor feedback would mean an improvement in his/her gestural performance. The computer's greater ability to recognize unknown gestural performances from a learner would seem to indicate that they are closer to the expert performances.

Figure 4.10 presents the confusion matrices of learner gesture recognition performed with and without computer-mediated feedback (11 learners and 5 repetitions, thus 55 instances). When no feedback was provided, the gesture with the lowest recall was the G3 because 17 repetitions of it were assigned to G2. This may be due to the fact that both G2 and G3 were performed within the same very limited spatial workspace (wheel-throwing diameter). The total precision and recall reached was 80%. Interestingly, when feedback was provided, there was a +11% in precision and +9% in recall. More precisely, the recall of G2 and the precision for G3 improved respectively from 67% and 77% to 100%. This improvement could be linked with the better spatial performance of gestures, as explained in the previous section. Most importantly, total precision and recall increased by approximately 10%, reaching approximately 90% for both metrics.

The analysis of learner's gestures, performed without feedback, permit-



	M1	M2	M3	M4	R (%)	TR (%)
G1	53	2	–	-	91	**80**
G2	8	42	4	1	67	
G3	–	17	35	3	65	
G4	1	–	–	54	98	
P	81%	70%	77%	93%		
TP	**80%**					

	M1	M2	M3	M4	R (%)	TR (%)
G1	53	1	–	1	96	**89**
G2	–	55	–	–	100	
G3	–	18	37	–	67	
G4	3	1	–	51	93	
P	95%	72%	100%	98%		
TP	**91%**					

Figure 4.10: Confusion matrix without (top) and with (bottom) feedback

ted the identification of temporal and spatial deviation in comparison to expert gestures. Motion capture, machine learning and gesture recognition allowed the similarity measurement of the learner's performance with the expert's, thus providing confirmation of the second hypothesis (H2). The reduction in temporal and spatial deviation of gestures performed with computer-mediated feedback assistance and the improvement of their recognition accuracy, validated the third and final hypothesis that sensori-motor feedback assists with self-training and contributes to improving a learner's performance.

4.9 Summary of contributions

In this chapter, a methodology for considering the computer as a partner in sensori-motor human learning was proposed and the concept was proved for the transmission of movement skills in the wheel-throw art of pottery.

In summary, three scientific contributions that justify the role of an HAI-driven computer as a partner in learning expert gestures were considered. The first contribution was the development perception layer for the computer that is able to recognize kinematic expert skills. The second contribution consisted of extending this perception layer in recognizing gestures executed by apprentices from a vocabulary of expert performances and measuring the distance between them. Finally, the third contribution offered validation of the hypothesis that computer-mediated sensori-motor feedback has a positive impact on the human self-training process.

Chapter 5

Digital musical instruments

Contents

5.1 The Big Picture

A musical instrument is a physical interface that can be considered as a means of musical expression and performance. The acoustic piano (pianoforte) is undoubtedly one of the most successful of human creations as it articulates intermediary mechanisms to trigger organised sounds. DMIs, and, in particular, MIDI keyboard instruments, are financially affordable, come in smaller sizes and offer great flexibility in terms of being able to select the synthesised sound. Although they are strongly inspired by the original sound-producing piano finger gestures, it is nevertheless a fact that they neither replace pianos nor use the whole potential of New Interfaces for Musical Expression (NIME). Mastery of both acoustic pianos and DMIs requires years of training, practice, and apprenticeship before players are able to perform. Thus, ‘learning’ musical gestures and ‘performing’ music are usually perceived as separate concepts and experiences, simply because there is no quick transition from novice to expert.

This chapter describes my scientific contribution to CCIs and more specifically to DMIs. Initially motivated by the research perspectives of my PhD in computer vision-based recognition of finger musical gestures [Manitsaris 2010], an HAI-driven DMI was designed and developed to address both learning and performing needs. The motion of the upper-body part occurring inside 2 bounding volumes above a tabletop is captured and sonified using explicit and implicit mapping strategies (Figure 5.1). In learning mode, the learner can investigate various piano-like fingerings to explicitly generate notes and sounds or implicitly modulate an audio excerpt from an expert performer through his/her gestures. In performance mode, finger and hand-based explicit instrumental mappings are proposed that make use of physically-based piano models and synthesize stringed or plucked-instrument characteristics. Moreover, accompanist gestures are also captured and used to control reverberation, stereo-panning and other sound morphologies.

The main motivation for this work was to investigate a new form of partnership between the human and the DMI, where the musical instrument would be much more than an interface, having as a unique goal the unidirectional control by the human over the sound [Chadabe 2002]. Motion sensing and HAI opened up completely new pathways in understanding the meaning of the musical gestures, thus proposing various forms of collaboration, mainly in performing and learning music. This collaborative process may unfold for years before a performance, or it may also happen to novices, making the creative process a co-discovery journey between the partners [Fiebrink 2017].

The scientific contribution presented in this chapter was conducted by supervising the PhD of Edgar Hemery on ’Modeling, recognition of finger gestures and upper-body movements for musical interaction design’ [Hemery 2017] and of Christina Volioti on ’Machine learning in sonification of expressive gesture with the use of stochastic models' [Volioti 2016], both funded by the i-Treasures FP7 project. A number of figures of this chapter are extracted from their theses.

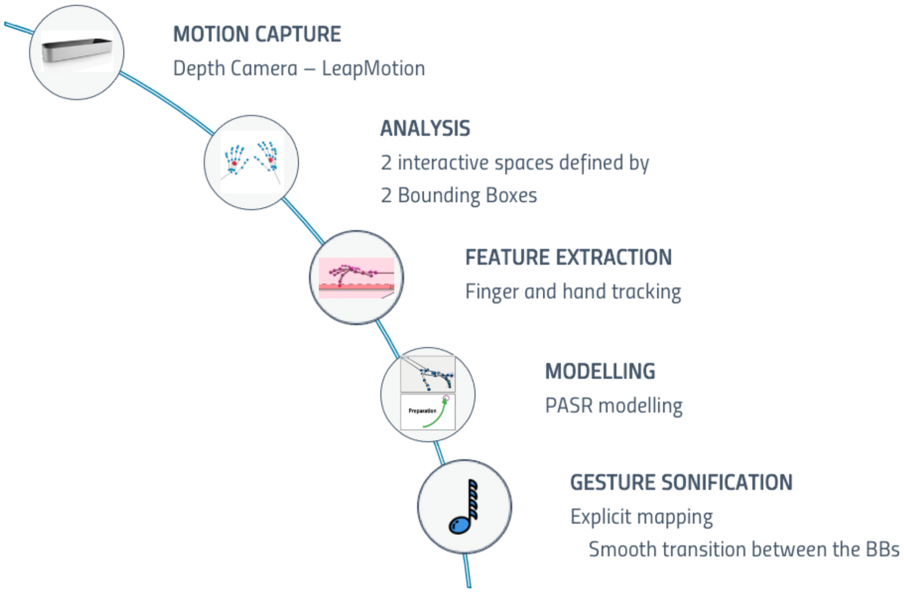


Figure 5.1: Generic methodology for gesture recognition and collaboration

5.2 State-of-the-Art

5.2.1 Typology and meaning of musical gestures in playing and learning

The gesture vocabulary described by Delalande, which has been used extensively in the literature [Delalande 1988, Cadoz & Wanderley 2001, Zhao & Badler 2001], categorises musical gestures into three classes: 1. *Effective gesture*: necessary to mechanically produce the sound (e.g. press a key); 2. *Accompanist gesture*: an auxiliary but non-sound-producing movement associated with the effective gesture; and 3. *Figurative gesture*: a gesture which is not related to any sound-producing movements, but which conveys a symbolic message.

In the in-person transmission of music-playing, the meaning of musical gestures varies according to 3 different perspectives: a) first-person; b) second-person; and c) third-person perspectives on gesture [Leman 2010].

The a) first-person perspective defines the meaning of the gesture for the person that actually implements it (whether expert/teacher or learner). An expert performer has developed the appropriate sensori-motor skills to be able to interpret a musical piece and, at the same time, the mental capacity to be able to translate the musical score into gestures and music, while a beginner has not.

The b) second-person perspective occurs during the in-person transmission (e.g. in music schools). First, a mirroring system is established between expert and learner: « my » perception of « your » gesture [Darwin 1872]. Then, the learner executes the gestures by translating previously observed expert gestures into 'concrete' actions.

The c) third-person perspective on gesture focuses on the measurement of moving objects, usually using audio-recording, video-recording, motion capture technologies and brain scans, as well as on physiological body changes [Camurri *et al.* 2005, Friberg & Sundberg 1999, Darwin 1872].

5.2.2 Gesture control of sound

In DMIs, gesture control of sound should be intuitive, thus translating motion data into sound in a reactive way. The 'mapping gesture to sound' (or 'gesture sonification') is the procedure where the motion data is being associated with the sound parameters. In explicit mapping, the input is directly associated to the output, while implicit mapping refers mostly to the use of machine learning techniques, which imply a training phase to set parameters [Françoise 2015]. For this reason, both the motion descriptors and the sound parameters that are going to be used for the sonification, need to be defined. In practice, the user performs a gesture that is recognized by the DMI and used to control the synthesized sound, following an explicit or implicit mapping strategy. As far as the recognition is concerned, the statistical and deep learning methods and approaches implemented have already been presented in Sections 2.2.2 and 2.2.3.

5.2.3 New Interfaces for Musical Expression (NIME)

Movement-based interactive systems, that allow for embodied performances through motion capturing and MIDI mapping, appeared in the 1990s. 'The Lady's Glove' was the first glove that transformed hand and finger gestures into sounds and it was developed for the purposes of the Ars Electronica Festival 1991 [Rodgers 2010].

Beyond the Digital Baton of the MIT Media Lab, various musical interfaces, such as the USB Virtual Maestro [Nakra *et al.* 2009], or the Modular Objects MO of IRCAM, made use of various inertial sensors or accelerometers to capture body motion and translate it into music.

Omnitouch wearable interfaces offer tangible interactions that are restricted to a flat surface with finger tapping, scroll, flick, pinch-to-zoom etc. [Harrison *et al.* 2011]. A well-known example of such interfaces is The ReacTable [Jordà *et al.* 2007]. It is a tangible tabletop for multi-player music interaction that uses real communicating objects. It is equipped with an infrared camera which recognizes objects on the table and activates sounds accordingly.

Nevertheless, more recently, motion data can also be obtained with computer vision and RGB-D cameras. The Leap Motion sensor opened new pathways for finger-based control of sound synthesis. For example, the GECO application [Hantrakul & Kaczmarek 2014] enables MIDI control through 3D mid-air gestures; Han and Gold [Han & Gold 2014] created an air-key piano and an air-pad drum; while the BigBang rubette controls notes, oscillators, modulators [Tormoen *et al.* 2014] etc. Finally, the Airpiano interface combined a Leap Motion with a transparent sheet of PVC to create a multi-touch musical keyboard for hovering control [d'Alessandro *et al.* 2015]. The Seaboard by Roli[1] and the TouchKeys[2] by Andrew McPherson are both extended piano keyboards which consider only fingertips.

5.3 Objectives beyond SoA

Leveraging on the above SoA, three objectives were defined:

O1: *Motion capturing finger and upper-body gestures in musical interaction*

Development of a markerless vision-based algorithm for capturing the motions of the whole upper-body in musical interaction.

O2: *Musical gesture recognition and sonification in 3D interactive sound spaces*

Development of an HAI-based perception layer that understands micro and macro human movements and sonifies them following various sonification strategies.

O3: *Digital Musical Instrument for performing and learning*

Development of a novel musical instrument which allows for piano-like music per-

[1] https://roli.com/products/seaboard

[2] https://touchkeys.co.uk

forming, learning and appreciation of music in general, without physical intermediary mechanisms.

The methodology developed for O1 contributes to the answering of Q1 (see Chapter 1, Section 1.2), by selecting the appropriate vision-based motion sensors and kinematic motion parameters for accurate recognition of upper-body gestures in musical interaction.

O2 contributes to the answering of Q2 (see Chapter 1, Section 1.2), by exploring the possibility of replacing current intermediary instrumental mechanisms for 'sensing' human motion and 'translating' them into sound by using HAI algorithms.

O3 contributes to the answering of both Q1 and Q2. With regards to Q1, it studies whether HAI algorithms can recognize not only sound-producing finger motions, but also accompanying upper-body movements. As far as Q2 is concerned, the main contribution concerns testing whether such a system of intelligence can be seen as a playful, embodied and expressive partnership between the human and the machine.

5.4 Overview of the Digital Musical Instrument

Our DMI should be regarded as work which contributes to the continuation of the NIME community. Inspired by the piano, we use in our system, metaphors of pianistic gesture. In past decades, as mentioned previously, MIDI keyboards became more and more popular because of their low price, their smaller size, bulk and weight and their flexibility in terms of sound choices. The downside of these instruments has been that they do not and cannot replace real pianos in terms of sound quality and they do not fully exploit the potential of what could be done with digital interfaces. With this in mind, the DMI brings the concept of keyboard instruments to the fields of gesture recognition and human-computer interfaces.

The DMI is an intuitive interactive system, which enables play with the fingers and upper part of the body. On a deeper level, it aims at capturing piano-like gestures in order to create sounds with them. These gestures are finally transformed into sounds via a 'mapping' phase; however, the objective is not a virtual replacement either of the piano or of any other keyboard instrument.

Moreover, the DMI can be used as a system of intelligence for musical pedagogy, where the learner interacts with it to master piano-like techniques or explore musical patterns of well-known composers. Although the interaction has been simplified for the purpose of having a smooth learning curve, it requires some practice in order to perform elements of musical stylistics like dynamics and articulation.

Many early and advanced prototypes of the DMI were developed mainly to explore various possibilities in the design and building of its 'table'. Nevertheless, all of them have a common purpose and base of sensors and algorithms. The first prototype of 70x40x13 cm was built using plexiglass (Figure 5.2) and was equipped with 2 Leap Motion sensors for detecting and tracking hand and finger positions in 3D space; 2 Animazoo IMUs mounted on the wrists for measuring of their 3D rotations; 1 Microsoft KINECT sensor for capturing the upper-body motions and obtaining 3D joint positions; and 1 Emotiv sensor to record electrical brain patterns (research partly sponsored by the i-Treasures partners). However, gestural interaction is not limited to this 2D surface, but instead extends to a volume of up to 30 cc above the plexiglass. The presentation and appearance of the DMI have progressively been transformed into a more compact and portable box (Figure 5.3) of 50x50x15 cm that can be more easily transported. The laptop(s) where the algorithms run can also be put on the lid of the box.

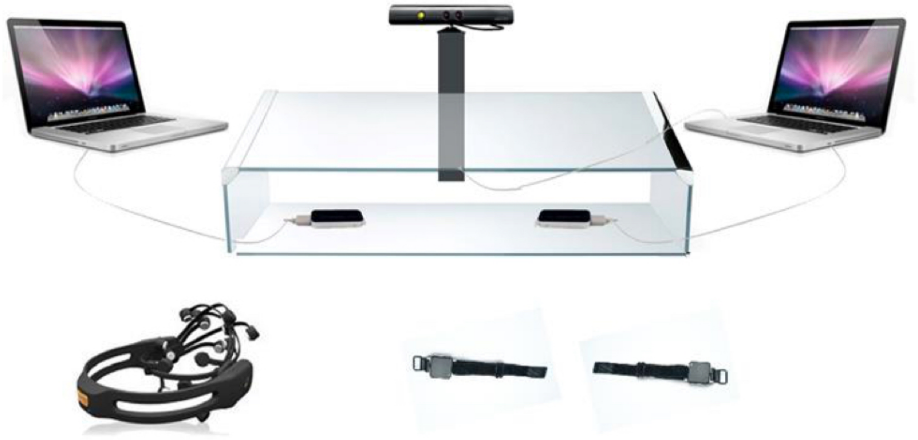


Figure 5.2: First prototype of the DMI

In general, creating the DMI (i.e. designing and implementing the architecture, the interaction design and the sound mapping software) has been a demanding process. We had to go back and forth between these three areas cyclically: eliminating skeleton joints, adding features, changing the tilt of the table, changing the material used for the table, experimenting with sound synthesis engines etc., until a coherent and satisfactory way to play the instrument was found.

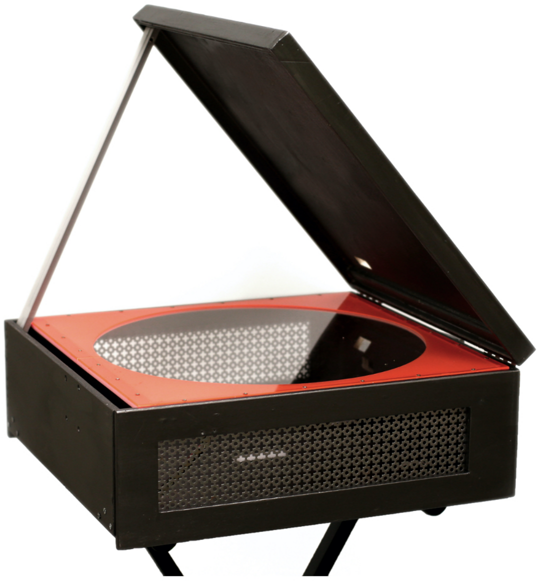


Figure 5.3: Latest prototype of the DMI, designed and implemented by Edgar Hemery

5.5 Bounding volumes for recognition and interaction

The movement-based interaction is performed inside two bounding volumes above the acrylic sheet (or plexiglass): the *micro* bounding volume, where finger and hand motions are detected, and the *macro* bounding volume, where body movements of a large amplitude are detected. The two bounding boxes play the role of 3D interactive spaces where gestures are recognised and sonified. A smooth/natural transition between them is supported.

The micro bounding box is a dematerialized bounding cone (pink cone in Figure 5.4), which is defined by the perpendicular axis to the Leap Motion and an angular span equal to the range of view of the sensor. Its intersection with the 'table' defines a circle with a 45 cm diameter. The depth of the cone depends on the lighting conditions and can go up to 45 cm above the surface. All micro interactions occur withing this bounding cone, and hands and fingers are not captured outside it.

Inside the bounding cone, a smaller bounding trapezoid, with a height of

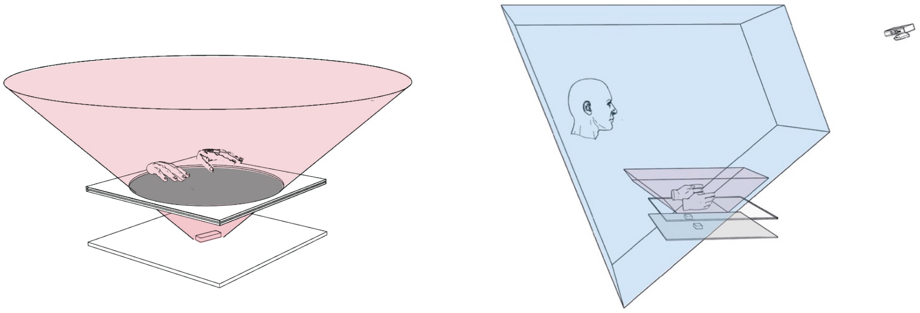


Figure 5.4: Micro (pink) and macro (blue) bounding boxes

1 cm above the acrylic surface is defined (red volume in Figure 5.5). Fingerings are only recognised inside the bounding trapezoid. In order to be able to calibrate the distance between the acrylic surface and the sensor, the user is asked to touch and slide on the surface with his/her fingertips for about 10 seconds. In this way,y-coordinates of the fingers are computed, thus providing a threshold, to which 1 cm is added, to obtain the position of the upper side of the trapezoid.

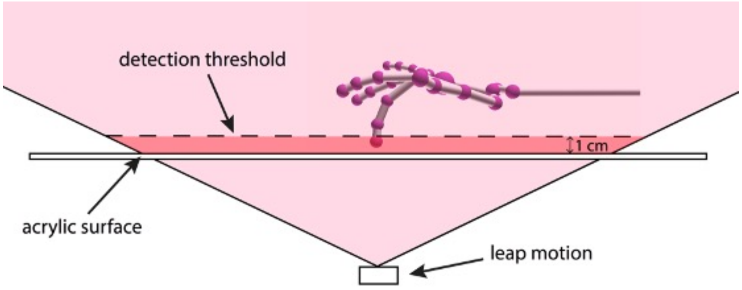


Figure 5.5: The y-axis detection zone (bounding cone) is in red and the Leap Motion field of view (bounding cone) is in pink

The *macro* bounding volume is defined by the Microsoft KINECT sensors that are placed approximately 1.2 m in front of and 1 m above the table (blue volume in Figure 5.4) and captures the head, right and left arms and torso.

The IMUs do not follow the principle of the bounding volumes and can operate independently and without the use of the cameras.

5.5.1 Modelling hand positions: the octave-like metaphor

In order for the DMI to recognize piano-like fingerings, the acrylic surface of the table has been segmented into a number of 'mental zones', which are represented by three different colors in Figure 5.6. Each zone consists of a set of fives notes, with each note corresponding to one fingertip. When a hand is in the zone, only one particular finger can play one particular note as its ID is associated with this particular note. For example, if the right index finger is inside the central zone, it will play the note 'D' or if the same finger is in the right zone, it would play the note 'G'. Thus, each hand covers three zones and two hands cover six zones, which correspond to 3 real octaves from C2 up to C5.

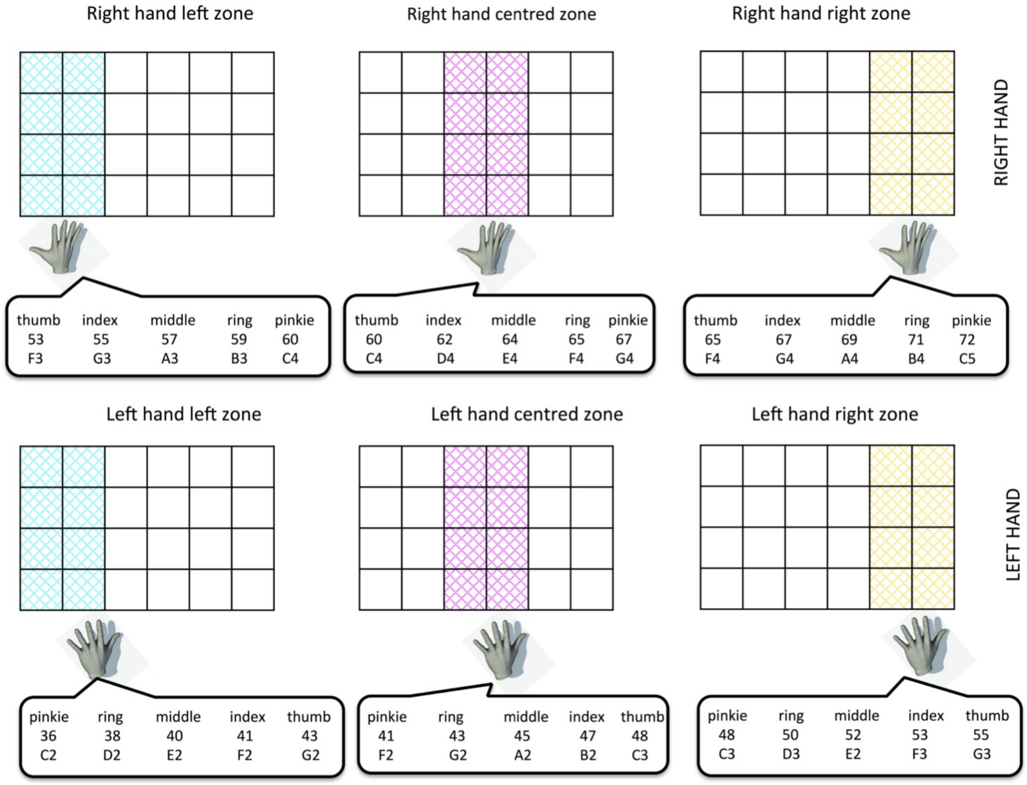


Figure 5.6: The keyboard metaphor: 3 zones per hand, 5 notes per zone

5.5.2 Modelling piano-like fingerings: the keyboard metaphor

Fingerings constitute the cornerstone of piano-playing and their retrieval on RGB sequences have presented a personal research challenge for many years [Manitsaris & Pekos 2008, Manitsaris & Pekos 2009, Tsagaris *et al.* 2011, Manitsaris *et al.* 2015, Manitsaris *et al.* 2016]. They are decomposed into a number of movement primitives that are internal states of fingering.

During the fingering, four movement primitives occur: preparation, attack, sustain, release (PASR). In the preparation state, one or several fingers lift upwards. In the attack state, one or several fingers go down and in sustain state (declination of the attack state), one or several fingers remain in contact with the acrylic surface. In rest position, the hand and fingers are relaxed and in contact with the table surface. The gestures terminate outside the bounding cone (Figure 5.7). Finally, a *release* state is considered as the initial and final state, which means that the process initiates and terminates when the fingers are outside the bounding trapezoid. Starting from the release state, the fingers can forward to any other states except to the rest position.

The motion descriptors that are extracted when a gesture occurs within the micro bounding volume are the 3D positions of the fingertips for both hands. Moreover, medium-level features are computed based on the 3D fingertips positions. These are 1. the *preparation time* that counts the duration a fingertip needs to go out of the bounding trapezoid and come back into it; 2. the *inter-onset-time* that counts the duration between two consecutive attacks; 3. the *sustain time*; and 4. the *rest time*.

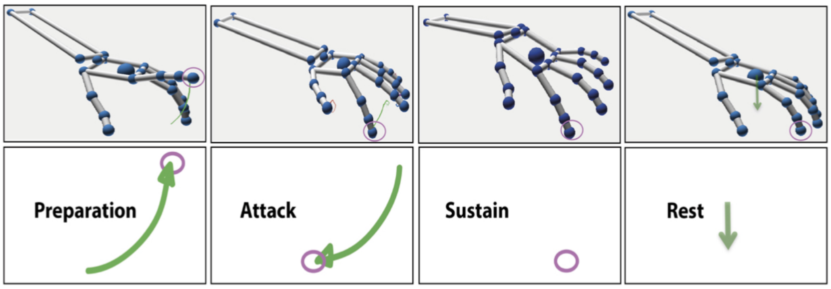


Figure 5.7: Movement primitives for fingering: preparation, attack, sustain and rest

5.5.3 The elastic rubber and kite-flying metaphors

Beyond the sound-producing piano-like gestures, a number of gestural metaphors have been integrated into the macro bounding volume. They are pantomimic gestures that convey a meaning by miming an action. The elastic rubber gesture is a metaphor where a rubber cable is stretched and released, which is based on the Euclidean distance between right and left hand. The kite-flying metaphor is based on a gesture deployed along a 3D rotation on a plane defined by positions of the head and the two hands.

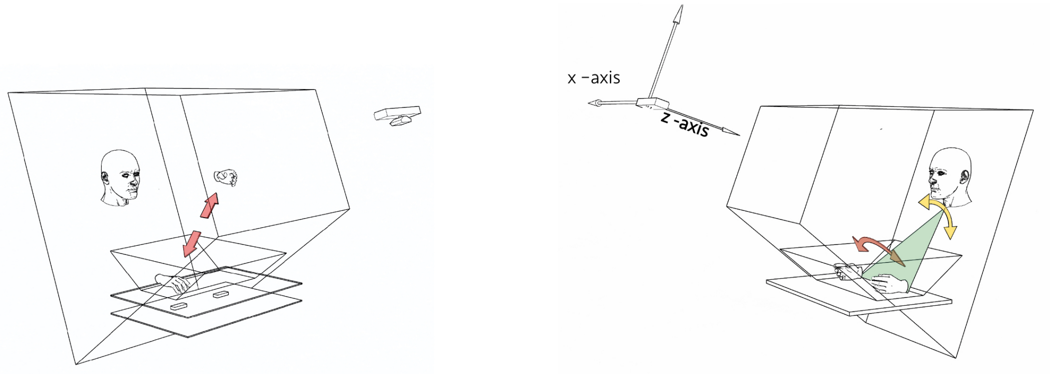


Figure 5.8: Left: The elastic rubber gesture with the lengthening/shortening of the Euclidean distance between hands. Right: The kite-flying gesture where the triangular plane reacts according to how much the human body is rotating on the left or on the right (z-axis - red narrow), going forwards/backwards, or hands are going up or down (x-axis - yellow narrow)

5.6 Gesture sonification strategies

5.6.1 Explicit sonification

In this section, a general overview of the most important explicit sonification (mapping) techniques that are integrated into the DMI is presented. For a more analytical presentation (e.g. classic fingerings in piano-playing vs cover fingerings on the DMI, how to play a complete musical scale on the DMI etc.), please refer to the PhD thesis of Edgar Hemery [Hemery 2017].

A form of instrumental mapping was implemented, within the micro bounding volume, which made use of both the keyboard and octave metaphor as well as the PASR model for the movement primitives. To accomplish this, the Pianoteq plugin, that is a physically-based piano model, was used [Rauhala *et al.* 2008]. The DMI used the note-on/note-off parameter to map fingering with a pitch/MIDI (frequency and duration), as well as the velocity parameter to map the velocity of the fingertip in the 'attack' state (intensity of the note).

Furthermore, for string sounds, a solution was found in the Karplus-Strong algorithm, as used in [Jaffe & Smith 1983], which proposes a system with 32 slightly de-tunable strings in parallel. In addition to the motion/sound mapped parameters for the piano sounds, the de-tune parameter was also used to facilitate a continuum of different parameters. Thus, the timbre continuum is attributed to the y-axis of the surface, while pitch is attributed to the x-axis, and velocity is attributed to the z-axis, as presented in Figure 5.9.

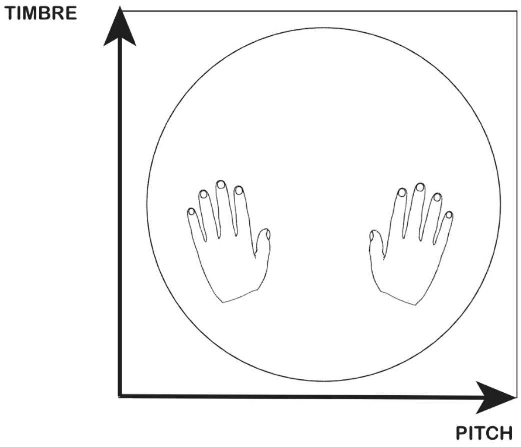


Figure 5.9: Pitch-timbre space

In addition, a filter control that maps the left-hand y-axis to the cutoff or center frequency was implemented. In practice, the filter is applied onto the sound that is triggered with the right hand, and its parameters are controlled, while the left hand goes up and down within the micro bounding volume.

5.6.2 Learning-by-demonstration for implicit gesture sonification

The implicit sonification strategy that was integrated into the DMI had a twofold goal: 1. *performing* while remaining confident as to the reference gesture provided by an expert; and 2. *learning* by imitation the musical gestures of an expert.

For this reason, the recognition and sonification engine x2Gesture was developed [Volioti *et al.* 2014, Volioti *et al.* 2015, Volioti *et al.* 2016a, Volioti *et al.* 2016b, Volioti *et al.* 2018], as it extends the GVF [Caramiaux *et al.* 2014, Zandt-Escobar *et al.* 2014]. GVF implements SS and Particle filtering for recognizing gestures using one-shot-learning. Moreover, GVF uses a pre-fixed value (*tolerance*) within the observation equation for defining the deviation between the various examples of the same gesture. The SS model of the GVF is based on the assumption that a gesture depends on a 3D variation space: *speed, scaling, rotation.* x2Gesture proposed an extension of the SS model of GVF by introducing the assumptions of the GOM, which have already been analytically presented in Section 2.4. Thus, the advantage of x2Gesture is that the deviation among the various observations of the same gesture is statistically estimated by the SS model and can dynamically change over time.

The algorithm includes two phases: the *learning* and the *following* phase. x2Gesture is first trained with a single expert example for each gesture, along with their attributed pre-recorded sound. In the phase that follows, the user imitates in real-time the same reference gesture. For each performed musical gesture, x2Gesture recognizes the input gesture and activates the appropriate confidence bounds. At the same time, the model aligns the incoming gesture onto the template gesture, estimating also the gesture variations. The system resynthesizes a plausible imitation of the original sound in real-time according to the user's gesture performance, which is described by the variables of speed, scaling, rotation and joint angles or positions. Then, a new audio signal is generated by using the granular sound synthesis engine. The better the recognition results are, then the better the gesture sonification and re-synthesis of the sound is.

In practice, by introducing the assumption of the GOM into the recognition engine, the expressive variations of the reference gestures (namely, 'expert gestures' since they are provided by the same user) are also taken into consideration through the confidence bounds and can dynamically change over time.

5.7 Fingering detection and explicit sonification assessment

The first assessment layer for the DMI consists of the evaluation of fingertip detection. To this end, the inter-onset metric for 3 dynamic ranges, namely those of p, mf and fff in an ascending C-Major scale, were used as a criterion. For this test, 10 successive performances of the same scale were recorded at the dynamic ranges of p, mf and fff. The typical fingerings for an ascending scale are thumb-index-middle-thumb-index-middle-ring-pinkie (8 fingerings). The results of the test are presented in Figure 5.10. Since no missing detection occurs, only false positives are presented. Accordingly, 8x10 fingerings per dynamic range were performed with 7.5% error for p, 1.25% error for mf, and no error for fff, while 80% of the detections were generated by the thumb. Since the intensity of the fingering is related to the velocity of the fingertip, it was observed that the lower the velocity of the fingering was, the more the error in the detection of the fingertip increased.

p series	false positive	*mf* series	false positive	*fff* series	false positive
*p*1	middle B$_5$	*mf*1	0	*fff*1	0
*p*2	0	*mf*2	0	*fff*2	0
*p*3	0	*mf*3	thumb D$_5$	*fff*3	0
*p*4	0	*mf*4	0	*fff*4	0
*p*5	index F$_5$, middle A$_5$	*mf*5	0	*fff*5	0
*p*6	0	*mf*6	0	*fff*6	0
*p*7	thumb A$_5$	*mf*7	0	*fff*7	0
*p*8	0	*mf*8	0	*fff*8	0
*p*9	thumb B$_5$	*mf*9	0	*fff*9	0
*p*10	thumb C$_6$	*mf*10	0	*fff*10	0
p error	**7.5%**	*mf* error	**1.25%**	*fff* error	**0.0%**

Figure 5.10: Fingering detection on ascending scale for p, mf, fff

In order to understand what the impact of the false positives was on the musical performance, a closer look at the velocities of the fingerings was taken; thus, it was observed that the velocities of the false positives were much lower than the velocities of correct fingerings. When the ratio between correct and false fingering was computed (Figure 5.11), it was concluded that the smaller the ratio, the larger the velocity and the intensity was. Thus, false detection fingering that triggered sounds with small intensities were masked by the correct note. In conclusion, only $p5$ and $p9$ false fingerings were intelligible because their ratio was underneath a threshold.

series	correct finger	false positive	velocity ratio
$p5$	thumb F$_5$	index, middle	0.623, 0.572
$p7$	middle A$_5$	thumb	0.158
$p9$	ring B$_5$	thumb	1.59
$p10$	pinkie C$_6$	thumb	0.140
$mf3$	index G$_5$	thumb	0.223

Figure 5.11: Velocity ratio between correct fingering detection and false positive

Linear mapping between fingertip velocity and sound intensity (loudness) was implemented. In order to test whether this principle was confirmed in practice, the loudness of a C-Major scale played 5 times for p, mf and fff, and thus 15 recordings in all, were recorded by averaging the maximum amplitudes of the sound over the last 5 milliseconds. The average peak was finally converted into decibels. Therefore, there was a difference of 7.91 dB between p and fff with a standard deviation of 1.92dB, mainly because of the imprecision in the calculation of the loudness (only 5 samples over 1024). Moreover, the linearity among the different range dynamics was confirmed, as shown in Figure 5.12.

Furthermore, the notion of latency is of a great importance for any musical instrument, whether acoustic or digital, because we cannot cognitively accept any intelligible delay between the effective gesture and sound production. As far as the DMI is concerned, two tests were conducted to test latency, where an expert performer was asked to produce: 1. a single note 10 times with the index fingertip; and 2. an *arpeggio* (thumb-index-middle-pinky) 4 times at different tempos, varying from *adagio* (80 Beats Per Minute (BPM)) to *prestissimo* (380 BPM). According to Figure 5.13, there is latency of 57ms for repeated notes and 51ms for the *arpeggio*.

5.8 Gesture recognition and implicit sonification assessment

During the development of the DMI, many functional and technical assessments of implicit sonification were executed. They are analytically presented in the PhD thesis of Christina Volioti [Volioti 2016]. Because of the synthetic goal of this thesis, 2 assessment case-studies are presented: 1. performing musical gestures using the

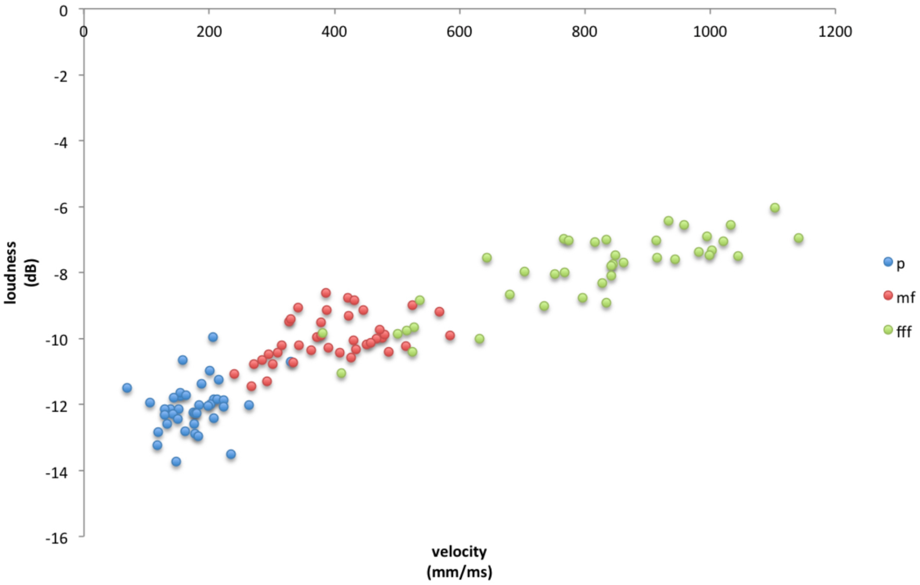


Figure 5.12: Linearity among the different range dynamics

DMI; and 2. learning-by-imitation expert musical gestures [Volioti *et al.* 2015].

5.8.1 Assessing the accuracy in performing music

The following musical gestures are included in the musical vocabulary: (a) G_1 ascending scale performed in legato style, (b) G_2 descending arpeggio performed in staccato style, and (c) G_3 a musical excerpt from a famous Greek song. The duration of the gestures is between 10 - 15 seconds and the 6 users repeated each gesture 5 times, following various tempos: adagio, andante, moderato, allegro (Figure 5.14). Inertial sensors were mounted on the 2 wrists of the users and 3D Euler angles were extracted. For both GF and GVF the tolerance was predefined at 0.1, while for x2Gesture it was dynamically attributed through the SS models for the 2 hands and their confidence bounds. The algorithms were trained with the template gestures and the pre-recorded sounds. In the recognition phase, x2Gesture selected the appropriate confidence bounds. In sonification, the sound was re-synthesized online, while the gesture was replayed, by using the granular synthesis engine. The metrics that are described in Section 2.9 were also used for this assessment. They were extracted following the jackknifing procedure and the gestures that were used

Single Note		Arpeggio	
Tempo	Avg (ms)	Tempo	Avg (ms)
80 bpm	60	60 bpm	44
120 bpm	57	70 bpm	51
180 bpm	56	90 bpm	56
220 bpm	55	110 bpm	49
260 bpm	54	130 bpm	55
300 bpm	55	140 bpm	51
340 bpm	57		
380 bpm	59		

Figure 5.13: Average latency according to BPM for repeated notes and arpeggio

for training and recognition came from the same user (mono-user test).

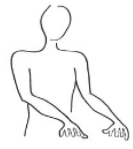
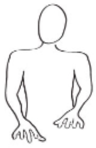
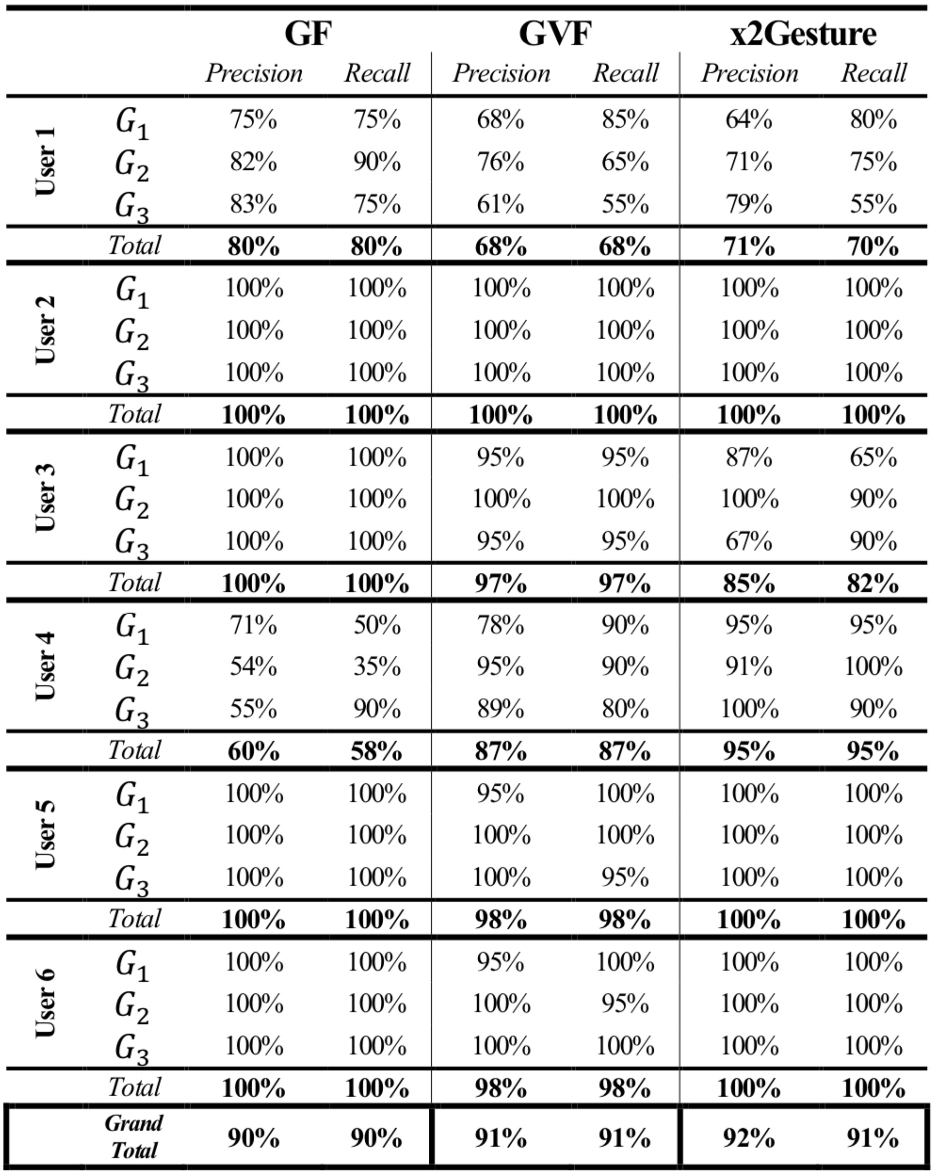
 (a)	 (b)	 (c)
Slow – 72 bps (adagio)	*Slow* – 80 bps (andante)	*Slow* – 72 bps (adagio)
Normal – 100 bps (andante)	*Normal* – 112 bps (moderato)	*Normal* – 100 bps (andante)
Fast – 116 bps (moderato)	*Fast* – 126 bps (allegro)	*Fast* – 116 bps (moderato)

Figure 5.14: (a) G_1 ascending scale, (b) G_2 descending arpeggio, (c) G_3 musical excerpt from a Greek song

The recognition results are shown in Figure 5.15 and, in general, x2Gesture outperforms the other algorithms by at least $+1\%$ precision while at user level GF gives better results for 4 out of the 6 users.

Nevertheless, the recognition performance of the algorithms is only one parameter, while the stability in recognition is also important. Generally, the less the algorithm 'hesitates' in recognition, the better it is for sonification, since each 'hesitation' generates a 'jump' to a different sound. As a result, according to Figure 5.16,



		GF		GVF		x2Gesture	
		Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
User 1	G_1	75%	75%	68%	85%	64%	80%
	G_2	82%	90%	76%	65%	71%	75%
	G_3	83%	75%	61%	55%	79%	55%
	Total	**80%**	**80%**	**68%**	**68%**	**71%**	**70%**
User 2	G_1	100%	100%	100%	100%	100%	100%
	G_2	100%	100%	100%	100%	100%	100%
	G_3	100%	100%	100%	100%	100%	100%
	Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
User 3	G_1	100%	100%	95%	95%	87%	65%
	G_2	100%	100%	100%	100%	100%	90%
	G_3	100%	100%	95%	95%	67%	90%
	Total	**100%**	**100%**	**97%**	**97%**	**85%**	**82%**
User 4	G_1	71%	50%	78%	90%	95%	95%
	G_2	54%	35%	95%	90%	91%	100%
	G_3	55%	90%	89%	80%	100%	90%
	Total	**60%**	**58%**	**87%**	**87%**	**95%**	**95%**
User 5	G_1	100%	100%	95%	100%	100%	100%
	G_2	100%	100%	100%	100%	100%	100%
	G_3	100%	100%	100%	95%	100%	100%
	Total	**100%**	**100%**	**98%**	**98%**	**100%**	**100%**
User 6	G_1	100%	100%	95%	100%	100%	100%
	G_2	100%	100%	100%	95%	100%	100%
	G_3	100%	100%	100%	100%	100%	100%
	Total	**100%**	**100%**	**98%**	**98%**	**100%**	**100%**
	Grand Total	**90%**	**90%**	**91%**	**91%**	**92%**	**91%**

Figure 5.15: Precision and Recall for GF, GVF and x2Gesture

x2Gesture is more stable during the recognition process and gives the appropriate maximum likelihood (Figure 5.17) faster than the other algorithms. In this case, all three algorithms correctly recognised G_3; however, automatically the sound synthesis with x2Gesture would be expected to be smoother, given that all 3 algorithms used the same sonification mechanism and sound synthesis engine.

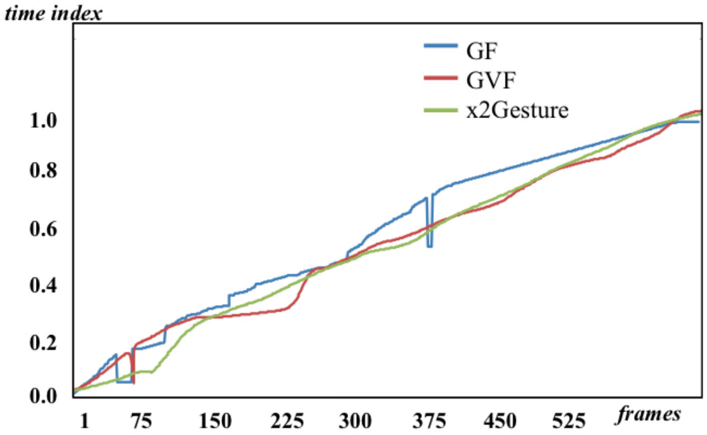


Figure 5.16: Temporal alignment and gesture progression for G_3 from User 3 from GF, GVF and x2Gesture

5.8.2 Assessing accuracy in learning music

In order to evaluate the use of the x2Gesture algorithm for learning musical gestures, 1 expert pianist was asked to perform the musical vocabulary and then 6 music apprentices were asked to observe his personal style and perform the same gestures. The musical dataset contained 90 examples (6 users x 3 gestures x 5 repetitions).

The data of the expert was used for training the algorithms while the data of the apprentices was used for recognition and synthesis. Figure 5.18 presents the result of the comparison for the 3 algorithms. They are consistent with what was expected since they confirm the hypothesis that the confidence bounds generated from the SS models of the expert can improve the recognition performance. An improvement of at least +7% for Precision and +4% for Recall was provided by the x2Gesture on a multi-user dataset.

5.8.3 Sound similarity between training and sonification

In order to test the hypothesis that better and smoother recognition means better sonification, a similarity measurement between the original and the re-synthesized sound was conducted.

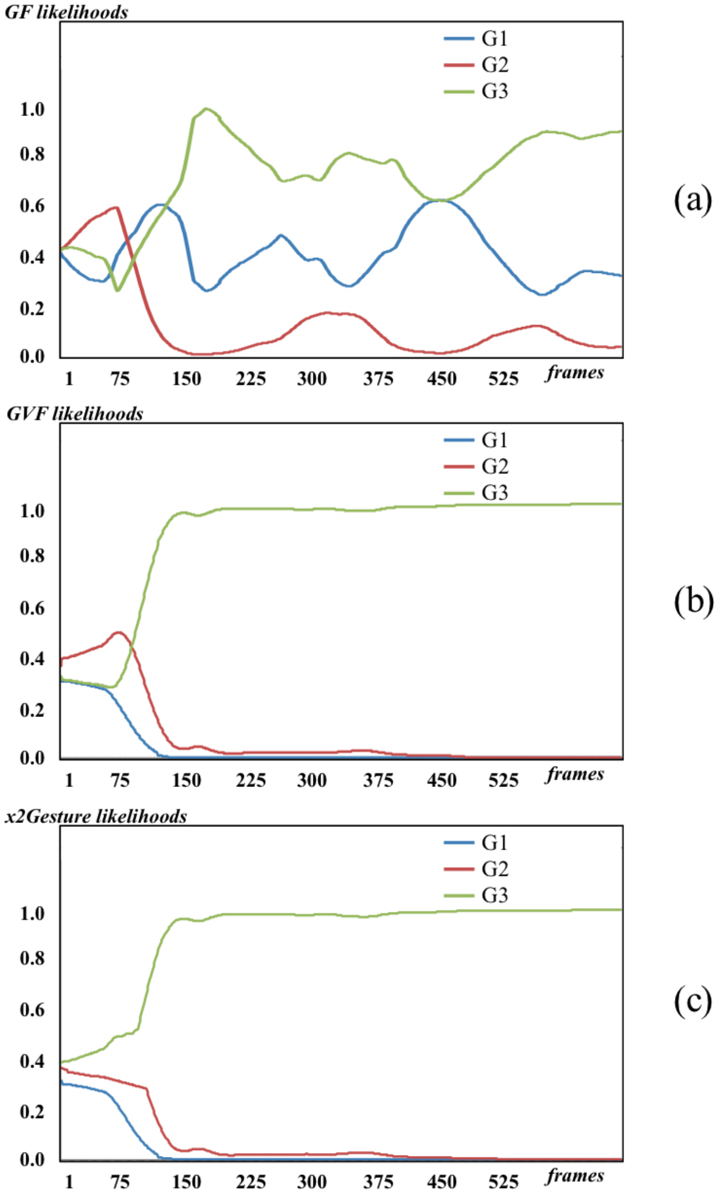
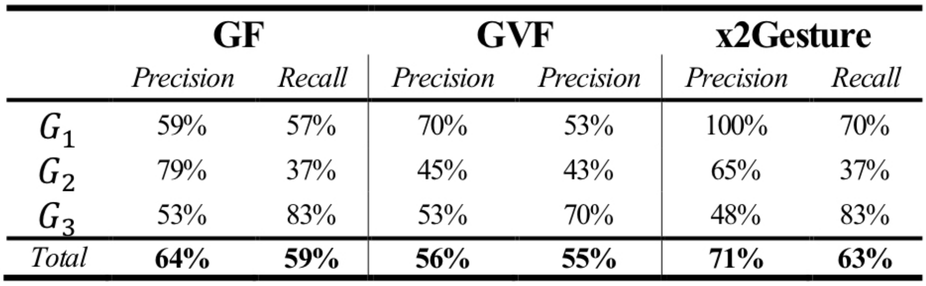


Figure 5.17: Instant likelihoods for G_3 from User 3 from GF, GVF and x2Gesture

In this instance, the procedure of [Tchernichovski *et al.* 2000] was followed. It consists of a Fast Fourier Transformation (FFT) that transforms the waveform into a frequency space. The FFT window on the original sound describes its characteristics (e.g. tonality, pitch, Wiener entropy etc.), which follow a different sta-



	GF		GVF		x2Gesture	
	Precision	*Recall*	*Precision*	*Precision*	*Precision*	*Recall*
G_1	59%	57%	70%	53%	100%	70%
G_2	79%	37%	45%	43%	65%	37%
G_3	53%	83%	53%	70%	48%	83%
Total	**64%**	**59%**	**56%**	**55%**	**71%**	**63%**

Figure 5.18: Precision and Recall on a multi-user dataset for GF, GVF and x2Gesture

tistical distribution. In order to obtain a global similarity between 2 sound signals, the sound characteristics are converted into statistical distances. For example, the tonality that is usually measured in Hz is converted to the median absolute deviation and the Euclidean distance is then computed on the new characteristics. For each pair of windows from the original and the re-synthesized that is labeled as 'similar', the probability that the 'match' would have occurred by chance is computed. The lower the probability is, the higher the similarity between the two samples. The final 'matching' is measured using the similarity metric, which counts the number of audio segments of the original sound that are contained in the re-synthesized one, as well as the accuracy, which generates the median local similarity for the whole signal and quantifies the 'imitation' ability of the new signal.

The similarity and accuracy for the musical vocabulary are provided below for the x2Gesture, GF and GVF algorithms. Beyond the fact that the gesture sonification from x2Gesture is slightly more similar to the original one and the generated sound is at least +1.7% more accurate than from GF or GVF, there is also a confirmation of the initial assumption which is that better recognition means better sonification. For example, GF and GVF have a delay in recognizing G_3 (Figure 5.17), which has a direct impact on the similarity between the sounds (Figure 5.19). Furthermore, slight delays in performing the gesture by the user may occur. They could also be characterised as an expressive variation in the performance. x2Gesture has the ability to remain correctly aligned with the appropriate template gesture and to continue sonifying less accurately, but without delays, mainly due to the use of the confidence bounds.

	GF		**GVF**		**x2Gesture**	
	Similarity	*Accuracy*	*Similarity*	*Accuracy*	*Similarity*	*Accuracy*
G_1	95%	76%	96%	76%	96%	77%
G_2	96%	76%	96%	77%	96%	78%
G_3	96%	78%	96%	79%	96%	81%
Total	**95.7%**	**76.7%**	**96%**	**77.3%**	**96%**	**79%**

Figure 5.19: Similarity and accuracy in sound synthesis compared to the original sample for GF, GVF and x2Gesture

5.9 Summary of contributions

In this chapter, a methodology for conceptualising, designing and developing a new DMI that is able to capture the whole upper-body motions, including fingers, and translate them into music without any physical intermediary mechanism was considered.

In short, two major scientific contributions emerged from this study. First, there was validation that in addition to effective musical finger gestures, other gestures can modulate or generate sounds when an HAI perception layer is added to the DMI. Equally well, there was evidence that HAI-driven DMIs can be musical partners since they can contribute to the performance, learning and appreciation of music in general, without using any physical intermediary mechanism.

Chapter 6

Synthesis and perspectives

Contents

6.1 Other contributions and ongoing work

6.1.1 Vocal tract imaging and speech synthesis for post-laryngectomy voice replacement

This work was done during my postdoctoral research at the ESPCI ParisTech. The aim was to develop a voice replacement technology that permits speech communication without vocalisation, also known as SSI. The visual-speech recognition engine used in the proposed SSI is based on vocal tract imaging [Denby *et al.* 2011, Cai *et al.* 2013]. The advantage of this engine is that it gives the opportunity to the individual who has undergone laryngectomy to speak using his/her original voice. Lip and tongue movements are recognized in real-time with the goal of extracting text sentences, which are synthesized afterwards by a Text-to-Speech (TTS) system. I contributed significantly to the creation of English and French voices, based on semi-HMMs [Manitsaris *et al.* 2012]. The SSI system is installed on a backpack and can be controlled remotely using a mobile device, and the new voices are installed on a Web Server. This work was funded by the "Revoix" ANR research project.

6.1.2 3D reconstruction of a deformable revolving object under heavy hand interaction

In this work, the goal was to reconstruct a 3D deformable object over time, in the context of the live creation of a ceramic vessel by an expert potter. The hands interact heavily with the clay in the creation process of the deformable object. Particle energy optimization was used to extract the profile of the object and use its radial symmetry to increase the quality of the reconstruction, thus removing occlusion effects and noise. One or more streams of depth images can be used as input to the algorithm [de Charette & Manitsaris 2019]. This work was funded by the i-Treasures FP7 European project and was primarily developed by Raoul de Charette, under my supervision.

6.1.3 3D hand and fingers pose estimation for gesture recognition

In this work, a 3D hand and fingers pose estimation model was built that is based on depth imaging and Random Decision Forests. A time-of-flight sensor was used to capture the finger motions. The algorithm was initially used to capture the finger musical gestures with a semi-closed palm [Dapogny *et al.* 2013]. Subsequently, the algorithm was improved to also cover parts of the upper body in driver-vehicle interaction. More precisely, two use-cases were considered for application inside the cockpit of a vehicle: 1. palm-grasp micro-gestures at the steering-wheel; and 2. macro gestures in front of the touchscreen. The goal of this research was to remove physical instrumental mechanisms (i.e. buttons) from the dashboard of a car in order for the driver to be able to keep his/her hands on the steering-wheel and to maintain his/her gaze on the road [Jacob *et al.* 2015, Pradere *et al.* 2019]. This work was funded by the PSA group of companies and implemented by Yannick Jacob, under my supervision.

6.1.4 Gesture control of mobile robots

The use of a 2D tablet to control a drone in a battlefield imposes strong limits when the soldier in control is constantly moving and yet needs to be focused on his/her task. We worked on the validation of the scientific hypothesis that air-gestures, as a modality to command a drone, place fewer constraints on human behaviour and demand a less significant cognitive load than the use of a tablet. First, an interaction model was designed to define the most ergonomically appropriate gesture vocabulary. Second, a gesture recognition module was developed that was based on a formal description of regular expressions. The experiments conducted confirmed the initial hypothesis that gestures are a less distracting modality, both visually and physically [Taralle *et al.* 2015a, Taralle *et al.* 2015b, Taralle 2016]. This research was conducted in the context of the PhD of Florent Taralle, under my supervision, and was funded by the SAFRAN group of companies.

6.1.5 Automatic detection of work-related musculoskeletal disorders

Manual, heuristic and subjective methods are currently applied to identify the risk of having Work-related Musculoskeletal Disorders (WMSD) (e.g. European Assessment Worksheet (EAWS), Rapid Upper Limb Assessment (RULA) etc.). I am currently supervising the PhD of Brenda Olivas, which aims at the analysis and forecasting of the exposure of the operators to postural risks, the identification of the body joints contributing the most to the movement and the automatic evaluation of the risk factors [Menychtas *et al.* 2019, Olivas *et al.* 2019, Olivas *et al.* 2020a, Olivas *et al.* 2020b, Menychtas *et al.* 2020]. Wearable sensors are used to record human motion and the time-series of joint angles produced are provided to the GOM. The impact of this work is twofold: 1. to characterize what is a 'good' and a 'bad' gesture from an ergonomic point of view; and 2. to use the output of this algorithm in order to delegate the most painful tasks to the machine in HRC configurations.

6.1.6 Egocentric gesture recognition using 3DCNNs for the spatiotemporal adaptation of collaborative robots

I am currently working on egocentric computer vision approaches for the recognition of actions and gestures of operators when collaborating with robots. The main goal of this research is twofold: 1. the development of an end-to-end DL architecture of 3DCNNs that receives egocentric images as input and outputs the actions of the operator assembling a TV screen; and 2. the quantification of the contribution of pose estimation and gesture recognition in HRC. Five experiments were conducted in which the following modalities for interaction are compared: 1. the operator presses on a button to activate the task of the robot; 2. as in (1) but the robot also estimates the pose of the operator so as to provide the parts at an ergonomic location; 3. as in (2) but with an additional sound notification to the human when the hand is waiting at a non-reachable location; 4. the recognition of the actions and gestures of the operator replaces the pressing of the button to activate the tasks of the robot and 5. the modalities for 2, 3 and 4 are put together (sound notification, pose estimation and gesture recognition). Furthermore, the quality of the collaboration was evaluated through questionnaires. The action recognition works almost perfectly, with an accuracy which is higher than 99%. The preliminary conclusions of this study show that the operators prefer to interact with a robot through gestures than by pressing a button and they consider the collaboration more natural. They consider the pose estimation useful for ergonomic purposes and the sound notification important for reminding them about an action [Papanagiotou *et al.* 2021]. Moreover, the collaboration that is based on action recognition is faster than the other modalities. A potential contribution of this work would be the proposal of a new Key Performance Indicator (KPI) that measures the impact of the spatial adaptation of the robot to the anthropometrics of the operator in HRC. More precisely, this KPI could be linked with the ratio of distance that an operator covers when the robot understands the human gestures to the covered distance when s/he presses a button to activate the task of the robot. This work is still ongoing and is being implemented by the research engineers Gavriela Senteri and Dimitris Papanagiotou. It is funded by the Collaborate H2020 project.

6.2 Discussion

In this thesis, overall I have presented an HAI approach for collaborating with machines through body motion. The principle of this approach relies on the creation of perception layers external to the machines that allow for sensing and understanding of human gestures (Chapter 2). Nevertheless, the scope of collaboration with the machines, or ‘systems of intelligence’, can be the goal, such as in robotics (Chapter 3), or the means, such as in human learning or in music playing (Chapters 4 and 5).

Indeed, it is possible for a machine to recognize kinematic parameters of situated expert and non-expert gestures (Q1 from Chapter 1). The research challenges for recognizing generic human actions from image sequences, such as challenges related with *inter*- and *intra*- class variations, lighting conditions, position and motion of the camera, limited access to annotated data or intention detection and predictability, are also valid for the recognition of situated gestures. Nevertheless, recording human motion in professional environments is a major challenge and data can be extremely rare, no matter what the sensors are. When the amount of data is sufficient and the goal is to classify human motion patterns, not many robust alternatives to deep learning exist. More precisely, when analysis, feature extraction, representation and modelling are not needed, end-to-end DL architectures constitute a valuable shortcut between motion sensing and recognition, especially for image data. However, recognizing a gesture may be much more complex than classifying it. Recognition can only be the final step of a bigger process where analysis, representation and modelling are also essential, such as for detecting not only *which* gesture is performed but also *how* it is performed. Therefore, decomposing the gesture into states for understanding its temporal evolution or concatenating the spatiotemporal dynamics of the body parts when they work synergistically, still constitute challenges, for which probabilistic methods and biomechanics may be able to provide valuable solutions. Finally, what applies in many scientific domains probably also applies in the case of gesture recognition: *the models should follow the data and not the opposite.*

In fact, it is possible to use gesture recognition as an alternative to instrumental interaction mechanisms (Q2 from Chapter 1). Instrumenting objects and tools with sensors for interaction purposes does not necessarily put limits on human behaviour. Indeed, when the data provided by pre-existing 'non-intelligent' tools (e.g. a standard screwdriver) fits in with 'systems of intelligence ', this generates new knowledge and means. In particular, when the machine is considered as a partner of the human, touchless interaction contributes towards a more natural collaboration. In HRC, gesture recognition facilitates the continuous monitoring and anticipation of the operator, as well as the 'on-the-fly' adaptation of the robot. In human learning, the computer becomes the partner of the apprentice when practising alone without the expert, due to the recognition of his/her trial-error gestures and the augmented feedback it provides. In musical interaction, gesture recognition and sonification enables embodied, real-time creation of the sound, removing physical constraints for performers and facilitating the learning curve for learners.

6.3 Achievements

Modelling the spatiotemporal dynamics of the human body for human action recognition and forecasting: My contribution in this field is the methodological representation and modelling of the spatiotemporal dynamics of the human body for gesture recognition and forecasting of motion trajectories. The operational model introduced for gestures relies on biomechanical principles to stochastically represent the temporal evolution of the dynamics of body movement. Furthermore, the human body is a physical model whose dynamics change over time and are influenced by endogenous and exogenous parameters. The concatenation of the dynamics of the theoretical model generates a set of first-order differential equations that fully describe the kinematics of the body movement. The model offers various forms of analysis for investigating which body parts contribute more to the gesture, over which axis, and how fast or slow, as well as what movement variations can be tolerated (confidence bounds). Various uses can also be considered, such as that of selecting the appropriate features and analysing the body dexterity or ergonomic risks related with the gesture itself. What we have proved is that, when this approach is used for gesture recognition, it outperforms state-of the-art probabilistic methods, as well as a standard end-to-end deep learning architecture.

Human-Robot Collaboration: My contribution to the science of HRC relies on the proof of the hypothesis that the professional gestures and actions of the operators can be continuously recognised in real-time and communicated to the robot. The robot can anticipate the human movements and adapt its behaviour accordingly. Furthermore, we proved that multi-user action and gesture recognition techniques can be deployed both in co-presence and collaborative workspaces. Both wearables and vision-based sensors can be used to capture human motion. The choice of sensor depends on the type of activity or industry and the constraints of the application. By instrumenting objects and tools with sensors, additional information about human actions can be recorded and used for improvements in the performance of the algorithm. We proved that when a new operator joins the line, only a few examples of his/her gestures are sufficient for the algorithm to recognize him/her properly. I consider my contribution to the field to be the innovation in the way the robot can be regarded in relation to humans. The robot becomes a partner of the human, opening up a huge gamut of perspectives and potential with regard to such collaboration, especially in improving the quality of life for humans at work.

Computer-mediated sensori-motor human learning: My contribution to this field relies on the premise that expert kinematic skills in manual jobs can be modelled and recognized using motion capturing and machine learning. A particular focus has been placed on the in-person transmission of skills in craftsmanship. The computer can be trained on some expert performances and a gamification mechanism generates augmented sensori-motor feedback when the apprentice deviates from the expert gestures. We proved, at least for the wheel-throwing art of pottery, that computer-mediated sensori-motor feedback has a positive impact on the human self-training process. I consider my main contribution in this area to be an innovation in the way expert movement skills can be recorded, preserved and transmitted to upcoming generations, both for industrial and cultural heritage purposes. We underline the role of the expert, who is irreplaceable, in the in-person transmission, and also prove the valuable role the computer has to play as a partner in the learning process.

Digital Musical Instruments: My main contribution to the field of DMIs and to the NIME community in general has been based on proving that it is possible to musically interact on tabletop 3D interactive spaces, while the whole body contributes in this creative process. Moreover, it is possible to go beyond the activation of physical intermediary mechanisms that trigger sounds through finger motions only. Inspired by piano playing, keyboard-like and octave-like metaphors are implemented. They generate a dynamic micro bounding box where mid-air finger movement primitives are detected and converted into physically-based piano and string sounds. Additionally, when hands are in the air, the upper-body motions are recognized, through stochastic machine learning, and are translated into sounds through a macro bounding box. All the gesture recognition and sonification mechanisms are integrated in a new tabletop DMI, which is able not only to recognize a number of musical gestures but, in addition, can also recognise some parameters of expressivity. Research and development has led to an innovation in the sense that we built a tabletop DMI, which can be a unique partner for humans, not just in the performing of music, but also in learning music.

6.4 Perspectives and challenges

Leveraging on the validation of the research questions Q1 and Q2 (Section 1.2), that are introduced in Chapter 1, and the quantitative evaluation of the methodologies proposed, I have provided concrete examples of HAI-driven movement-based human-machine collaborations. In fact, the research methodology of my thesis is industry-oriented and has produced prototypes for the Factory of the Future (mainly manufacturing), the Creative and Cultural Industries, as well as for Vocational Training. I also had the opportunity to perform various feasibility studies, or to supervise research, in the domains of interaction with intelligent vehicles (automotive industry) or drones (defense and security industries).

Although I have not had the opportunity to confirm my results beyond the industrial sectors previously mentioned, I am generally convinced that machines can be trained to recognize situated gestures and actions as well as anticipate them during a collaborative task. I therefore consider *generalisation* and *extrapolation* as my future (big) personal research challenges for gesture recognition.

Generalisation is definitely a well-identified problem in machine learning. From a human movement perspective, recognizing motion patterns from a relatively small vocabulary is possible. Nevertheless, because of the nature of humans and the variations derived in performing a movement, generalisation might be difficult, even with gestures of the same difficulty and nature e.g. professional gestures in the same job. Therefore, the recognition and characterisation of motion patterns from a large number of classes is not a straightforward issue. Again, if the classification of human actions is the only goal, the most recent advances in DL can provide solutions when sufficient data is available. Therefore, my principal motivation is to further investigate how machines can learn satisfactorily well a large number of patterns from the same group e.g. almost all the gestures involved in one profession. However, recognition is usually neither the final step in the pipeline nor the only output of the algorithm. Recognition and characterisation strongly depend on the application. For example, learning a movement from an expert performance or adapting a movement already learned to external disturbances are discrete challenges in sensori-motor human learning. They both need models that are capable of continuously decoding micro (adaptation to an existing skill) and macro (acquisition of a new skill) kinematic variations in the execution of the movement. Furthermore, differentiating a tolerated expert performance from a non-tolerated one is still a research challenge. Bearing this in mind, GAIL can be considered to be a promising research direction due to its capability to discriminate expert demonstrations from artificial non-tolerated ones.

Extrapolation is also a well-identified problem in AI that consists of being trained on a certain type of data and being capable of predicting another type of data. From the perspective of human movement, extrapolation would mean training the machine on a very large set of movement primitives that are common across various professional sectors e.g. screw, wait, put, rotate etc., with the goal of hierarchically recognizing and characterising the human activity as a sequence of actions constituted from movement primitives. For this reason, meta-learning can be a promising research direction due to its facility of *learning how to learn*, thus extrapolating from a set of actions to unseen actions.

I consider that in only a few years' time, almost all machines will be equipped with various layers of HAI-driven perception, enabling continuous understanding of behaviours and anticipation of human actions. Therefore, from a research perspective, *generalisation* and *extrapolation* are progressively becoming a necessity in order to address the future vertical 'lift-off' of demand for algorithms that are able to support human-machine collaborations on a grand scale.

Appendices

Chapter 7

Appendix with a few relevant publications

sensors

Article

Stochastic-Biomechanic Modeling and Recognition of Human Movement Primitives, in Industry, Using Wearables

Brenda Elizabeth Olivas-Padilla *, Sotiris Manitsaris , Dimitrios Menychtas and Alina Glushkova

Centre for Robotics, MINES ParisTech, PSL Université, 75006 Paris, France;
sotiris.manitsaris@mines-paristech.fr (S.M.); dimitrios.menychtas@mines-paristech.fr (D.M.);
alina.glushkova@mines-paristech.fr (A.G.)
* Correspondence: brenda.olivas@mines-paristech.fr

Abstract: In industry, ergonomists apply heuristic methods to determine workers' exposure to ergonomic risks; however, current methods are limited to evaluating postures or measuring the duration and frequency of professional tasks. The work described here aims to deepen ergonomic analysis by using joint angles computed from inertial sensors to model the dynamics of professional movements and the collaboration between joints. This work is based on the hypothesis that with these models, it is possible to forecast workers' posture and identify the joints contributing to the motion, which can later be used for ergonomic risk prevention. The modeling was based on the Gesture Operational Model, which uses autoregressive models to learn the dynamics of the joints by assuming associations between them. Euler angles were used for training to avoid forecasting errors such as bone stretching and invalid skeleton configurations, which commonly occur with models trained with joint positions. The statistical significance of the assumptions of each model was computed to determine the joints most involved in the movements. The forecasting performance of the models was evaluated, and the selection of joints was validated, by achieving a high gesture recognition performance. Finally, a sensitivity analysis was conducted to investigate the response of the system to disturbances and their effect on the posture.

Keywords: movement modeling; state-space representation; gesture recognition; wearable sensors; ergonomics

check for updates

Citation: Olivas-Padilla, B.E.; Manitsaris, S.; Menychtas, D.; Glushkova, A. Stochastic-Biomechanic Modeling and Recognition of Human Movement Primitives, in Industry, Using Wearables. *Sensors* **2021**, *21*, 2497. https://doi.org/10.3390/s21072497

Academic Editor: Zdeněk Svoboda

Received: 15 January 2021
Accepted: 26 March 2021
Published: 3 April 2021

1. Introduction

To fulfill market demands within specific time limits, job specifications and budget restrictions, the tasks performed by manual laborers in the industrial sector are becoming more challenging and complex. The tasks demanded of them require workers to go sometimes beyond their natural physical limitations, performing repetitive tasks for long periods of time. Being subjected to such constant physical strain leads to work-related musculoskeletal disorders (WMSDs) [1]. WMSDs can cause permanent or temporary damage to tissue, such as muscles, bones, joints or tendons, caused by cumulative microdamage, where the internal tolerance of the tissues is eventually exceeded. WMSDs are the most common work-related health issue in Europe [2], entailing consequences for workers and for the companies that employ them, that have to contend with high levels of sick leave and drops in productivity.

The ability to record accurate measurements for ergonomic analysis is essential as it provides ergonomists with quantitative measures of workers' performance. This represents an added value in preventing ergonomic risk. Risk factors such as assuming awkward postures and performing highly repetitive or physically demanding tasks are often associated with WMSDs [2], mostly when occurring at high levels of repetition or in some kind of combination. Several rules and methods were established to identify ergonomic risks that workers might be exposed to during their professional activities.

Three different measurements were used for these evaluations [3]. The first was self-assessment, where workers were asked to fill out a questionnaire indicating their level of exposure to diverse risk factors, including how tired they felt after their shift or if they had assumed any dangerous postures during their tasks. The second measurement is through observation by others, where an ergonomist observes the workers during their shift and completes a heuristic evaluation based on standards that indicate human physical limitations and abilities (e.g., the ISO 11226:2000 and EN 1005-4). These standards are mostly based on the deviation of the working posture from the neutral pose. The higher the deviation, the higher the risk of developing WMSD. Some existing questionnaries that use this approach are the Rapid Upper Limb Assessment (RULA) [4], Ergonomic Assessment Worksheet (EAWS) [5], and Ovako Working Posture Analysing System (OWAS) [6]. The third technique consists of direct measurement and primarily involves implementing a biomechanical-based analysis, where the loads and external forces the workers are exposed to are considered in the evaluation. An example of a direct measurement method is the National Institute of Occupational Safety and Health (NIOSH) lifting equation [7], which helps assess whether lifting a load is acceptable. Another is the Liberty Mutual manual materials handling tables [8], which indicate the load range that certain male or female members of the population may be able to lift, lower, carry, push or pull as part of their daily work, without the risk of developing WMSDs.

While methods based on self-assessment or visual observation, are quick and straightforward ways to evaluate, they are not always accurate and precise. They are quite subjective, since they are dependant on the worker's feelings or sensations, or on the powers of observation of the ergonomist, leading, quite possibly, to low accuracy and high intra- and inter-observer variability [9]. For methods based on direct measurements, laboratory equipment is usually required, such as optical motion capture systems and force plates to measure external forces. This equipment requires a large infrastructure and is thus rather impractical and difficult to use in the workplace. Moreover, using these technologies involves bringing workers to the laboratory, causing inaccurate measures since they lack authenticity and are not real workplace scenarios. Recent research has started to develop alternative sensor-based automated evaluation methods, using cameras or body-mounted inertial sensors [10–13]; however, ergonomic evaluation in these studies relies purely on joint angle thresholds, which can only identify risks related to static postures [4–6].

The work presented in this paper aims to further expand the scope of the analysis conducted in current ergonomic evaluations by modeling professional movements. The hypothesis formulated here is that by modeling the workers' dynamics, it is possible to extract information about the contribution of body joint movements to various ergonomic risks. Moreover, with the learned models, it is possible to predict the motion trajectory of body joints and thus detect any possible future exposure to postural ergonomic risk.

For the purposes of this research, human motion modeling and trajectory predictions were made using a Gesture Operational Model (GOM) [14], which consists of a system of equations based on different assumptions about the dynamic relationship of body parts. The methodology was validated by evaluating the forecasting performance of the system and by improving the recognition performance of professional movements, using four datasets. The first and second datasets were taken from professional movements executed in factories concerned with television production and airplane manufacturing, respectively. The third dataset was composed of gestures performed in a glassblowing workshop, while the fourth dataset of motion primitives, with different ergonomic risk levels, according to EAWS [5], was recorded in a laboratory.

In Section 2, which follows, the present state-of-the-art related to motion analysis for modeling, prediction, and pattern recognition, will be presented, while the methodology and evaluation procedures we used are described in Section 3. Section 4 presents the results of the experiments conducted on the four datasets, Section 5 discusses our findings and results, followed by the presentation of our conclusions in Section 6.

2. State-of-the-Art

2.1. Motion Analysis Based on Body Structure

In the past, biomechanical, stochastic, and hybrid models have been used to represent human motion and these models were then used to study the coordinated mechanical interaction between bones, muscles, and joints within the musculoskeletal system. The modeling of human movements, and their changes, caused by internal and external action forces has generally been addressed with biomechanical models. These models represent the human body as a set of articulated links in a kinetic chain where joint torques and forces are calculated using anthropometric, postural, and hand load data [15]. Inertial data, such as accelerations and velocities, and information about external forces like ground reaction forces from force plates, are used as input for biomechanical models [16]. When dealing with inverse dynamics, quantitative information about the mechanics of the musculoskeletal system, while performing a motor task, is extracted. Most previous studies have used biomechanical modeling to extract the kinematic and kinetic contributions of the joints, in diverse motor tasks, then investigate the mechanical loading of the joints and their response to ergonomic interventions. To analyze the ergonomic impact of different postures on human joints, Menychtas et al. [17] applied the Newton–Euler algorithm for the computation of upper body joint torques. The normalized integral of joint angles and joint torques was then calculated to describe the kinematic and kinetic contribution of the body joints when awkward poses are assumed. The method identified which joints moved the most during the tasks and were under the most strain while performing ergonomically dangerous gestures. Faber [18] used a spanned inverse dynamics model to estimate 3D L5/S1 moments and ground forces, then compared symmetric, asymmetric, and fast trunk bending movements through ergonomic analysis. Similarly, Shojaei [19] estimated the reaction forces and moments of the lower back, in manual material handling (MMH) tasks, to assess age-related differences in trunk kinematics and mechanical demands on the lower back.

In previous research, statistical modeling has been used to learn the stochastic behavior of human motion. These models capture the variance information of body motion trajectories and have been used both to estimate human intentions and label human activities. In order to infer intentions from observed human movements in real-time, Wang [20] presented the Intention-Driven Dynamics Model (IDDM), based on Gaussian processes. The dynamics model assumes that the goal directs human action, meaning that the dynamics change when the actions are based on different intentions. The study proved that including human dynamics in the modeling benefits the prediction of human intentions. In order to capture the motion patterns that emerge in typical human activities (e.g., walking and running), Argwal [21] trained a mixture of Gaussian auto-regressive processes with joint angles and position trajectories. The dynamic models take advantage of local correlations between joints motion to track complicated movements successfully (turns in different directions) using only 2D body measures (joint positions and joint angles). To segment and analyze human behaviours, Devanne [22] applied a Dynamic Naive Bayes model to capture the dynamics of elementary motions and to segment continuously in long sequences diverse human behaviors.

Hybrid methodologies that take into consideration human biomechanical structure and the stochastics of motion have been developed to improve the analysis of the random outcomes of movement. A hybrid model, designed to predict the probability of injury and identify factors contributing to the risk of non-contact anterior cruciate ligament (ACL) injuries, has been proposed by Lin [23]. A biomechanical model of the ACL estimated the lower leg kinematics and kinetics. In turn, the means and standard deviations of the number of simulated non-contact ACL injuries, injury rate, and female-to-male injury rate were calculated in Monte Carlo simulations of non-contact ACL injury and non-injury trials. *T*-tests revealed the biomechanical characteristics of the simulated injury trials. Donnell [24] used a two-state Markov chain model to represent the survival of surgical repair from rotator cuff. The load applied to the shoulder and the structural capacity of tissue were the random variables. The analysis was based on the application of structural reliability modeling. By introducing this new modeling paradigm for explaining clinical retear data, the model successfully predicted the probability of rotator cuff repair retears and contributed to understanding their causes. To describe the cooperation of body parts in the execution of professional movements, Manitsaris [14] proposed the Gesture Operational Model (GOM), based on state-space modeling. GOM offered insights into the dynamic relationship between body parts, within the execution of a movement, according to the statistical significance of its various assumptions and their dependencies on the motion of other body parts.

2.2. Motion Trajectory Prediction

The problem of human motion trajectory prediction has been researched extensively in the past. There is a growing interest, in the industrial sector, in implementing systems that allow prediction of how workers' motion descriptors will unfold over time, and to incorporate this knowledge in a pro-active manner e.g., to facilitate human-robot collaboration or risk prevention. There are three prediction approaches, which are based on how human motion is represented and how the behavior pattern is formulated. Physics-based models are explicitly defined dynamic models explicitly defined and follow Newton's Law of Motion. Pattern-based models, on the other hand, learn statistical behavioral patterns that emerge, based on the observed motion trajectories. Plan-based models are concerned with reasoning about the intention behind the movement and the goal of the performer.

2.2.1. Physics-Based Models

Physics-based models predict future human motions according to a defined dynamic model (f). This model follows the form of a state-space representation:

$$s(t+1) = f(s(t), u(t), t) + w(t) \tag{1}$$

where $s(t+1)$ is the prediction, $s(t)$ is the current motion state of the system, $u(t)$ is the input, and $w(t)$ the process noise. The motion is predicted by forward simulating the dynamic equations that follow the physics-based model. Physics-based models have tended to use kinematic models for prediction and these represented the motion states as position, orientation, velocity, or acceleration and linked the observations to the state's evolution. Some examples of kinematic models used are constant velocity (CV) [25], constant acceleration (CA) [26], and coordinated turn (CT) [27]. These models describe the agent's motion based on the mathematical relationship between the movement parameters (e.g., position, velocity, acceleration) without considering the external forces that affect the motion. Kinematic models are frequently used for prediction due to their simplicity and acceptable performance, under the conditions of little motion uncertainty, or short-term prediction.

For the prediction of pedestrians' position trajectories, previous studies have applied Kalman Filters (KFs), with kinematic models such as CV and CA [26,28]. The main application of KF is for tracking the pedestrian position according to the estimated velocity or acceleration. Zernetsch [29] applied a kinematic model for trajectory prediction of cyclists that consisted of a CV model for the computation of all significant forces, such as the driving force and resisting force, composed of acceleration resistance, rolling resistance, and air resistance. In order to determine the kinematic model parameters, a curve-fitting approach was used, with motion profiles of cyclists that were recorded with a video camera and laser scanners at a public intersection.

For the prediction of movements with a high level of uncertainty, previous studies have used multi-model (MM) methods. These methods fuse different motion modes (e.g., sudden accelerations, linear movements, maneuvers) to describe complex motions (e.g., pedestrians or vehicles in public areas), where a dynamic model represents each mode. Pool [30] applied an MM approach to predict cyclists' motion based on their motion strategies (go straight, turn left or right 45° or 90°). Whenever a strategy does not comply with the road topology, the probability of the strategy is set to zero, in place of prediction. A multi-model approach for pedestrian trajectory prediction has been presented by Kooij [31], which uses Switching Linear Dynamical Systems (SLDS) to model maneuvering pedestrians that shift between motion models (e.g., walking, stopping). Then, a Dynamic Bayesian Network (DBN) predicts the pedestrian movements based on the SLDS model. The latent variables consisted of the pedestrian location, curb location and head orientation (indicating awareness of oncoming vehicles). The results proved that including context cues in the analysis improves overall prediction accuracy. Manitsaris [14] adequately addressed the forecasting trajectories of a 3D skeleton's joint positions by using state-space modeling. The state variables corresponded with the dynamic association of body parts, their synergies, their serial and non-serial mediations, and the two previous positions of the body part represented. This study, by including information about other body parts in the representation of each body part, boosted the forecasting performance of the system due to the strong dynamic relationship between them.

Physics-based approaches are appropriate, where an explicit transition function can be defined for modeling the agent's motion dynamics, as well as the influence of other agents and of their surroundings on it. The main drawback of using physics-based approaches is that they do not perform well for very complex situations (e.g., public areas with multiple agents). Moreover, their use is commonly limited to short-term predictions and obstacle-free environments.

2.2.2. Pattern-Based Models

Pattern-based approaches, unlike physics-based approaches, learn human motion behaviors by fitting models to data. For the prediction of pedestrian trajectories, Quintero [32] presented the Gaussian process dynamical models (B-GPDMs). The system can reduce the 3-D time-related information extracted from key positions on the pedestrians' bodies into only two observations, used for the prediction. The most similar model to the multiple models of four activity types (e.g., walking, stopping, starting and standing) is then selected to estimate future pedestrian states. For the motion prediction of multiple people, Kucner [33] used Gaussian Processes and their mixtures to model multimodal distributions, representing speed and orientation in joint space, for the purpose of modeling the motion of people and mapping their flow in the area analyzed.

Neural Networks have achieved promising performances for time-series prediction [34–36]. Among the most popular are the Long Short-Term Memory (LSTM) networks to predict human [34,35] and vehicle motion [36]. For the trajectory prediction of pedestrians' 2D position and orientation, Sun [34] incorporated spatial and temporal context information into an LSTM to learn the human activity patterns generated in different environments at different times of the day. Xue [35] proposed the Social-Scene-LSTM (SS-LSTM), which uses three LSTMs to capture person, social and scene scale information. In turn, the output of the three networks is used by an LSTM decoder for the prediction of pedestrian trajectory coordinates. Srikanth [36] has proposed a robust model for future trajectory prediction of vehicles, where a simple Encoder-Decoder model connected by a convolutional LSTM was used to learn vehicle temporal dynamics, including semantic images, depth information and other vehicles' positions. In this study, the use of scene semantics improved the prediction performance over models that only use information such as raw pixel intensities or depth information.

For the capturing of more complex unknown dynamics, it has to be admitted that pattern-based approaches have outperformed physics-based approaches; however, they require a large amount of data to train the model to avoid generalization issues. To improve the prediction performance, pattern-based and physics-based approaches have benefited from integrating context information into their observations. The studies that included information about the shape and structure of the environment, together with the external forces that the person or object is exposed to, or information about their interaction with other agents (e.g., people, vehicles or robots) produce more precise predictions in numerous cases [37].

2.2.3. Planning-Based Models

The third prediction approach employs Planning-based models. Unlike the previous approaches, these assume rationality, in the case of tracked human movements and their long-term motion goals. This approach computes path hypotheses that allow the agent to reach their motion goals by considering the impact of current actions on future motions. The prediction is made using a predefined cost function, based on intended motion goals or inferred cost function, according to the observed trajectories. Best and Fitch [38] have proposed a Bayesian framework to estimate pedestrians' intended goal destination and future trajectory. The framework is based on multimodal hypotheses of the intended goal, and the long-term trajectory that decreases the distance to the intended goal is selected. By seeing the trajectory prediction as an optimization problem, Lee [39] suggests a deep stochastic Recurrent Neural Network (RNN) Encoder-Decoder framework for trajectory prediction of multiple vehicles in complex scenes. The model obtains a diverse set of hypothetical trajectories which takes into consideration the agent interactions, scene semantics, and expected reward function. The single end-to-end RNN encoder-decoder network captures the past trajectories and incorporates the information into the inference process to improve prediction accuracy.

In order to use planning-based approaches, the goals that the agents under analysis are trying to achieve must first be explicitly defined, and the context information about the environment surrounding the agent must be provided for the model. Planning-based approaches usually perform better for long-term predictions than do physics-based approaches and also tend to have less generalization issues than Pattern-based approaches. The downside of these approaches is that as the complexity of the prediction problem increases (e.g., long-term predictions, multiple agents and size of the environment), so does the running time for training the models.

2.3. Human Gesture Recognition

Ergonomic evaluations have been conducted by identifying the risks involved in work-related motions, using gesture recognition (GR) techniques to recognize professional motions and estimate their frequency and duration on the workers' shift. Peppoloni [40] developed a monitoring system by training State machines to classify manual handling activities with data from a wearable sensor network. Likewise, Ryu [41] trained a Support Vector Machine (SVM) classifier, with data from an accelerometer placed on the wrist, to classify a mason's actions (e.g., laying and adjusting bricks). With deep learning architectures, Slaton [42] trained a hybrid network, containing convolutional and recurrent Long Short-Term Memory (LSTM) layers, to recognize construction-related activities. Parsa [43] applied Temporal Convolutional Networks (TCNs) to segment videos and recognize manual handling tasks with different ergonomic risk levels.

Hidden Markov Models (HMMs) have been widely used for the modeling and recognition of human gestures. HMMs model the dynamic behavior of gestural time series based on a probabilistic interpretation of the gesture samples. The HMMs assume that a hidden state sequence causes the observed sequence (gesture samples). HMMs capture the motion patterns presented in the training set's gestures, meaning that they will not recognize other variations from these patterns that could emerge during the movement performance, after the training. To address this issue, Caramiaux [44] proposed the Gesture Variation Follower (GVF), representing pre-recorded template gestures with continuous state-space models. Particle Filtering was used to update the models' parameters to estimate the likeliest template of a new observation, considering its varying gesture characteristics. The gesture's speed, size, scaling and rotation angles were considered the varying gesture characteristics and state variables.

Despite the fact that ergonomic evaluation based on GR adds factors such as the frequency and duration of activities into the analysis, basing the ergonomic evaluation on only these two factors could lead to the oversight of other risk factors in the motions that could cause the development of WMSDs.

3. Methodology

Due to the nature of the hypothesis defined for this study, a physics-based approach was selected to model the dynamics of professional movements. Physics-based approaches have proved to be capable of handling joint predictions efficiently and because of the use of a transition function, they perform well with observations obtained from different environments and subjects, without extensive training datasets. This generalization capability is essential if workers from various industrial sectors are to be monitored. Moreover, by using a physics-based model, information could be extracted regarding the human dynamics and their response to risk factors, by examining the resulting trained models.

In this study, human motion was represented as a sequence of human poses, where each pose was described through 3D-joint angles. The modeling of each gesture was done using the Gesture Operational Model methodology [14], which was extended by integrating more assumptions into the representation of the motion of joints. The models were used to predict the trajectory of joint angles, instead of joint positions, to avoid forecasting errors such as bone stretching and invalid skeleton configurations, errors that commonly occur in models trained with joint positions [45–47]. The proposed methodology is illustrated in Figure 1.

The statistical significance of the assumptions of each model was computed to determine the body joints contributing the most to the professional movements. The selected joint angles were validated by comparing their gesture recognition performance with another two sensor configurations, the first using all joint angles for training, and the second using only a small set of two hand-picked sensors. Finally, the forecasting ability of the models was evaluated, and a sensitivity analysis was conducted to analyze the stability and behavior of the system when external forces affect system response, meaning a change in the posture and ergonomic risk level of the motion.

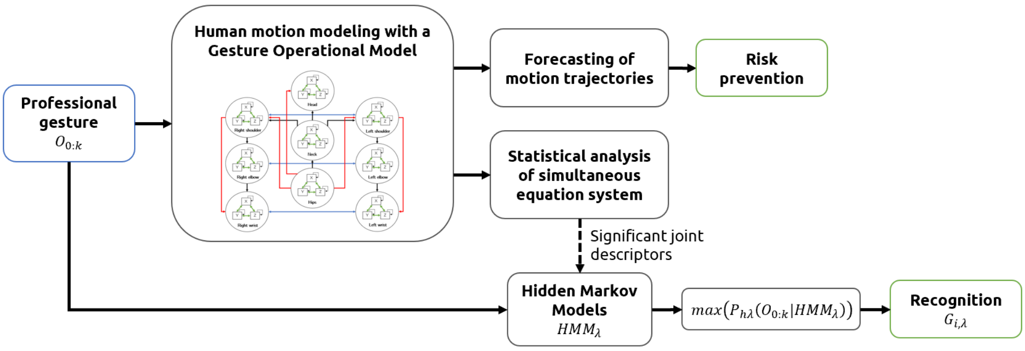

Human motion modeling with a Gesture Operational Model
Head
Right shoulder
Neck
Left shoulder
Right elbow
Left elbow
Hips
Right wrist
Left wrist
X
Y
Z
Professional gesture
$O_{0:k}$
Forecasting of motion trajectories
Risk prevention
Statistical analysis of simultaneous equation system
Significant joint descriptors
Hidden Markov Models
HMM_{λ}
$max(P_{h\lambda}(O_{0:k}|HMM_{\lambda}))$
Recognition
$G_{i,\lambda}$

Figure 1. Methodology pipeline.

3.1. Data Collection and Gesture Vocabularies

3.1.1. Inertial Motion Capture Technology

Due to the advantages of using motion capture (MoCap) technologies, based on inertial sensors for the MoCap of industrial workers and the subjects' movements, the BioMed bundle motion capture system from Nansense Inc. (Baranger Studios, Los Angeles, CA, USA) was used. This system consisted of a full-body suit composed of 52 IMUs placed throughout the body and hands. The sensors allowed the orientation and acceleration of body segments on the articulated spine chain, shoulders, limbs and fingertips to be measured at a rate of 90 frames per second. Those 52 rotations were combined to create a kinematic skeleton that included the body segments measured. The Euler local joint angles on three axes X, Y, and Z were computed through the inverse kinematics solver provided by Nansense Studio (suit software). The joint angles per time frame were then exported to Biovision Hierarchy (BVH) files. Before the analysis, an offline pre-processing procedure of the data was followed. The motion data was low pass filtered to mitigate noise, and the common zero velocity update algorithm was applied to remove the drifting caused by electromagnetic interference.

3.1.2. Recording and Gesture Vocabularies

Industrial workers from television (TV) production, airplane manufacturing, and glassblowing sectors were recorded under real conditions in their respective factories. Figures 2–5 illustrate the four gestures vocabularies, and a detailed description of each gesture is provided in the Appendix A.

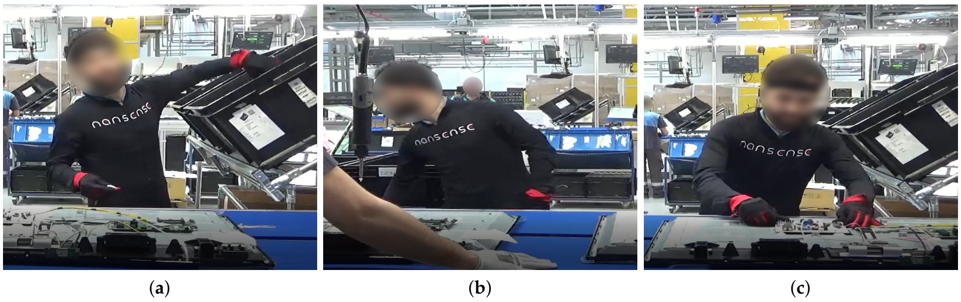

(**a**) (**b**) (**c**)

Figure 2. Gesture vocabulary with gestures for TV assembly (G_1). (**a**) $G_{1,1}$: Grab the electronic card from a container; (**b**) $G_{1,2}$: Take a wire from a container; (**c**) $G_{1,3}$: Connect the electronic card and wire and place them on the TV chassis.

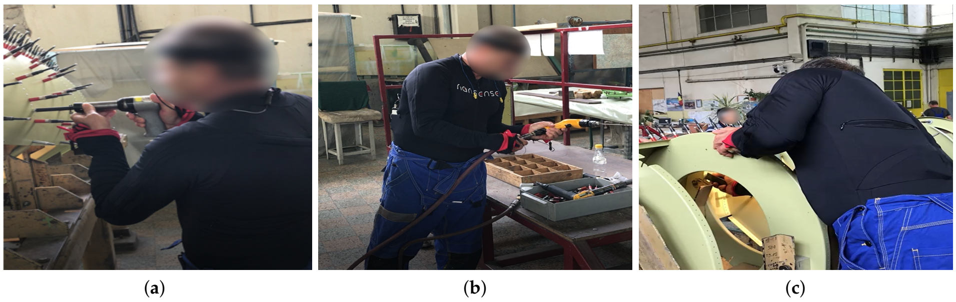

(**a**) (**b**) (**c**)

Figure 3. Gesture vocabulary with gestures for airplane assembly (G_2). (**a**) $G_{2,1}$: Rivet with the pneumatic hammer; (**b**) $G_{2,2}$: Prepare the pneumatic hammer and grab rivets; (**c**) $G_{2,3}$: Place the bucking bar to counteract the incoming rivet.

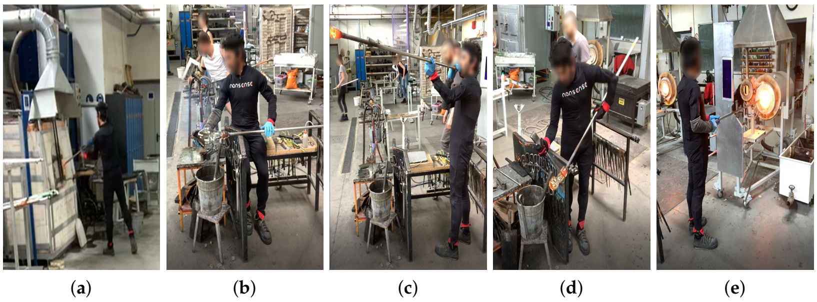

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 4. Gesture vocabulary with gestures for glassblowing (G_3). (**a**) $G_{3,1}$: Grab glass melt from the oven; (**b**) $G_{3,2}$: Shape the carafe's curves; (**c**) $G_{3,3}$: Blow through the blowpipe; (**d**) $G_{3,4}$: Shape the carafe's neck with pliers; (**e**) $G_{3,5}$: Heat the glass of the carafe.

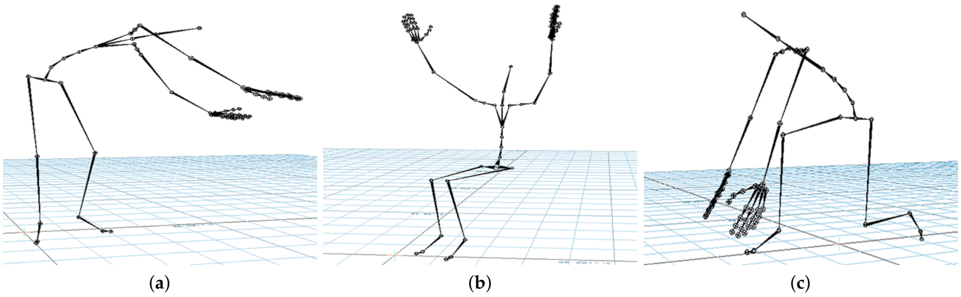


Figure 5. Gesture vocabulary with motion primitives based on EAWS (G_4). (**a**) $G_{4,1}$: Standing while bending forward and rotating the torso; (**b**) $G_{4,2}$: Sitting while raising arms above shoulder level; (**c**) $G_{4,3}$: Kneeling while bending forward.

The professional gestures from each gesture vocabulary presented essential differences in their execution due to the different contexts in which they were recorded. For instance, $G_{1,1}$, $G_{1,2}$, or $G_{2,2}$ were mostly manipulating tools or objects, where the subject grabbed an object or prepared it for later use. The iterations for these gestures had a high intra-class variance since their motion was not restricted, nor was high precision or dexterity required. On the contrary, for the gesture vocabulary G_3 and gestures $G_{1,3}$, $G_{2,1}$, and $G_{2,3}$, the subjects needed to be more precise since they placed the objects in a specific position. It has to be considered that human factors such as level of experience, fatigue, or mental stress affected how the subjects' bodies performed the gestures. Although this did not apply for the G_3, high dexterity and technicity were required to execute the gestures effectively. The gestures from G_3 were recorded from a glassblowing expert, who performed the gestures with high repeatability and low spatial and temporal variations between iterations, to produce a carafe four times with the same specifications. Regarding G_4, there was a low intra-class variation in this dataset since the performance of each movement was controlled. Feedback was provided to subjects in order to perform the gesture demanded correctly. On the other hand, the inter-class variation was intended to be low, where there were only a few variations in the postures assumed in each gesture. The end in using this last dataset was to test whether the proposed methodology was able to identify the small variations between motions and provide an accurate estimate of the joints that most contributed to the execution of the 28 motions.

From an ergonomic point of view, these four datasets could assist in the evaluation of human motions in industrial settings. The modeling of these motions could help in evaluating the subjects' manual dexterity in relation to the gesture's ergonomic risk. For example, as mentioned, the glassblower had a high level of dexterity for glassblowing gestures, but some observational methods could recognize that the gestures executed were ergonomically risky (e.g., $G_{3,3}$ and $G_{3,4}$). An ergonomic analysis of these gesture vocabularies could therefore aid in improving ergonomically how the professional gestures were executed without affecting manual dexterity.

3.2. Movement Representation with GOM

The Gesture Operational Model was composed of auto-regressive models that learn the dynamics of each body part. Each representation had different assumptions of the dynamic association between body parts. These assumptions consisedt of the intra-joint association (H1), inter-limb synergies (H2), serial (H3.1) and non-serial intra-limb mediations (H3.2), and transitioning over time (H4), [14]. For the intra-joint association, a bidirectional relationship was assumed between variables where the motion is decomposed e.g., joint angles on the X-axis, Y-axis, and Z-axis. The transitioning assumption was that current values depend on their previous values. The inter-limb synergies assumed a relationship between body parts that worked together to achieve a motion trajectory e.g., using both hands to execute a specific gesture. Finally, the serial and non-serial intra-limb mediations included the relationship between joints, whether directly and not directly connected e.g., the wrist was directly connected with the elbow (serial mediation) and indirectly connected with the shoulder (non-serial mediation). These assumptions are represented in Figure 6.

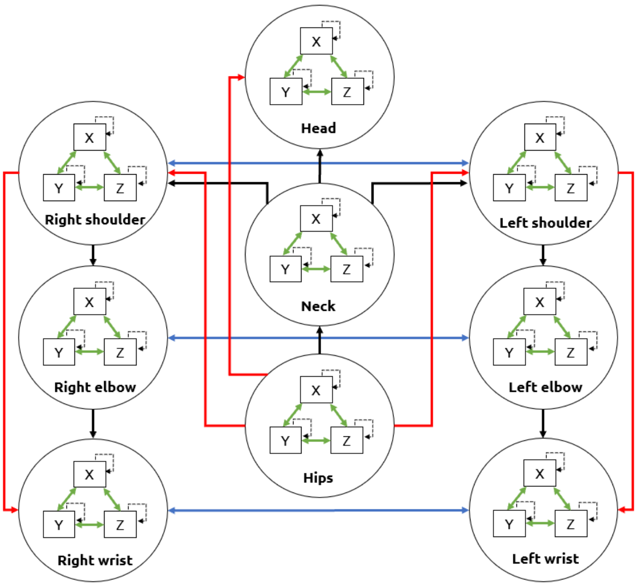


Figure 6. Upper-body assumptions that constitute a Gesture Operational Model. The intra-joint association is indicated by green arrows, transitioning over time with dashed arrows, inter-limb synergies with blue arrows, intra-limb serial mediation with black arrows, and intra-limb non-serial mediation with red arrows.

The number of representations was equal to the number of associated dimensions for a given body part, multiplied by the number of body parts defined in the GOM. The representation of each body part had different assumptions depending on its location within the body:

- Intra-joint association: All body parts included it in their representation.
- Inter-limb synergies: Only the body parts representing joint angles from arm and leg parts included this assumption.
- Intra-limb serial and non-serial mediation: The assumptions included in each representation depended on the body part location within the body. The joint angles related to the spine only included in their equation joint angles of other spine parts with which it had serial or non-serial mediation. The serial and non-serial mediations from angles related to the spine are illustrated in Figure 7a. The angles related to the arms only included in their equation joint angles of other arm parts with which it had serial or non-serial mediation (Figure 7b). Equally, the angles related to the legs only included in their equation joint angles of other leg parts (Figure 7c).

The transitioning assumptions corresponded to the lagged endogenous variables, where lag depended on the order given to the model. For this work, second-order autoregressive models were selected. The order was selected according to the correlation between lag values in the time series (auto-correlation). If the observations had positive auto-correlations with a certain number of lags, then it was better to have a higher order of differencing until the auto-correlation was negative and more than -0.5, to avoid overdifferencing [48].

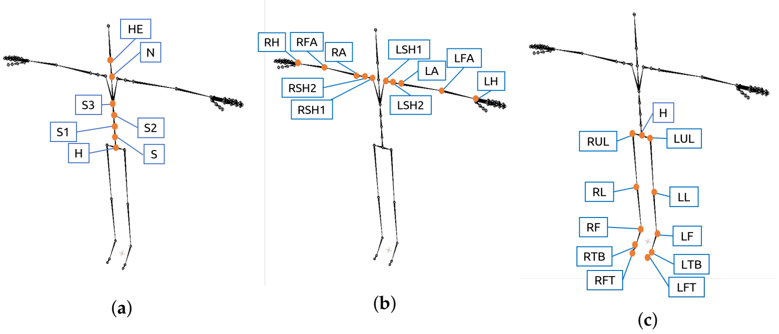


Figure 7. Location of the sensors that provide the XYZ joint angles included in GOM's state-space equations. (**a**) Spine parts; (**b**) Arm parts; (**c**) Leg parts.

An example of a mathematical representation of the assumptions is shown in Equation (2), for a motion on the X-axis (X_{ax}) of a body part P, with only two dimensions X_{ax} and Y_{ax}, and assumptions that includes an association with only a second body part ($P_{2,X_{ax}}(t-1)$).

$$P_{1,X_{ax}}(t) = \underbrace{P_{1,Y_{ax}}(t-1)}_{\text{H1}} + \underbrace{P_{2,X_{ax}}(t-1)}_{\text{H2}} + \underbrace{P_{1,X_{ax}}(t-1)}_{\text{H4}} + \underbrace{P_{1,X_{ax}}(t-2)}_{\text{H4}} \tag{2}$$

These representations were then translated into simultaneous equations by using state-space modeling. State-space equations allowed estimation of the state of the system according to the input-output data [49]. Thus, given the input and the current state of the system, state-space gave the hidden states that resulted in the observable variables. A state-space representation is shown in Equations (3) and (4). Equation (3) is the state-space equation, a first-order Markov process where A is the transition matrix. Equation (4) is the measurement equation, where the time derivative of the state vector $s(t)$ is taken into account for the computation of the output $y(t)$ along with the input vector $u(t)$, where C is the output matrix and D the feed-through matrix.

$$s(t) = AS_s(t-1) + w(t) \tag{3}$$

$$y(t) = Cs(t) + Du(t) \tag{4}$$

To model the GOM representation of the Equation (2) using second-order state-space modeling, first, the state-space variable is substituted with the subtraction of two previous values of the body part to model, each multiplied by one coefficient of the transition matrix:

$$s(t) = AS_s(t-1) = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} P_{1,X_{ax}}(t-1) \\ -P_{1,X_{ax}}(t-2) \end{bmatrix} = \begin{bmatrix} \alpha_1 P_{1,X_{ax}}(t-1) \\ -\alpha_2 P_{1,X_{ax}}(t-2) \end{bmatrix} \tag{5}$$

For the measurement equation, the input vector $u(t)$ corresponds to the endogenous variables, for the case of Equation (2), it consists of the intra-joint association and inter-limb synergy:

$$P_{1,X_{ax}}(t) = \begin{bmatrix} 1 & 1 \end{bmatrix} s(t) + \alpha_3 P_{1,Y_{ax}}(t-1) + \alpha_4 P_{2,X_{ax}}(t-1) \tag{6}$$

Finally, by merging Equations (5) and (6), the state-space representation is obtained:

$$\begin{aligned} P_{1,X_{ax}}(t) = \alpha_1 P_{1,X_{ax}}(t-1) - \alpha_2 P_{1,X_{ax}}(t-2) + \\ \alpha_3 P_{1,Y_{ax}}(t-1) + \alpha_4 P_{2,X_{ax}}(t-1) \end{aligned} \tag{7}$$

The full body modeling consisted of three sets of equations for each body part, one for each dimension X, Y, and Z. Hence, by discarding the body parts from the fingers, the GOM consisted of 84 equations per gesture. The coefficients of the equation system were estimated using the Maximum Likelihood Estimation (MLE) via Kalman filtering [50].

For the coefficient estimation, first, the probability of obtaining the observation vectors $O_{0:k}$ was defined:

$$P(O_{0:k}) = \prod_{t=0}^{k} P(O_t|O_{0:t-1}) \tag{8}$$

which consisted of the products of probabilities of the observation at time t, given previous observations. This probability distribution is considered Gaussian, as shown in the following equation:

$$P(O_{0:k}|\psi) = \prod_{t=1}^{k} \exp\left\{ -\frac{\left(o_t - \tilde{o}_t^{t-1}\right)^2}{2F_t^{t-1}} \right\} \left(2\pi\left|F_t^{t-1}\right|\right)^{-\frac{1}{2}} \mathrm{d}' \tag{9}$$

where F_t^{t-1} is the covariance and $\widetilde{o}_t^{t-1}$ is the mean. From Equation (9), the log-likelihood was computed, where the Kalman filter could optimally estimate the mean and covariance that gave the maximum likelihood:

$$\log L(\psi|O_{0:t-1}) = -\frac{k}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{k} \log\left|F_t^{t-1}\right| - \frac{1}{2}\sum_{t=1}^{k} \frac{\left(o_t - \tilde{o}_t^{t-1}\right)^2}{F_t^{t-1}} \tag{10}$$

The Kalman filtering consisted of two steps which were repeated until obtaining the maximum likelihood. These are known as the prediction and update steps. Initial values were set, then the log-likelihood was computed for the evaluation in the prediction step. Next, in the update step, the variance and mean were updated according to the Kalman gain (K_t), until, in the prediction step, the maximum likelihood was achieved:

$$K_t = \frac{F_t^{t-1}}{\left(F_t^{t-1} + R\right)} \tag{11}$$

$$\tilde{o}_t^{t-1} = \tilde{o}_t^{t-1} + K_t\left(o_t - \tilde{o}_t^{t-1}\right) \qquad F_t^{t-1} = F_t^{t-1} - K_t F_t^{t-1} \tag{12}$$

In the end, the computation of the coefficients of the state space models was derived through Equation (10).

3.3. *Applications of the GOM*

3.3.1. Selection of Significant Joint Angles

Statistical analysis was done to investigate the significance of the model assumptions in relation to the body part associations defined within the GOM. By estimating the statistical significance of each assumption, it was possible to determine which joint descriptors contributed the most to the execution of all the gestures of each gesture vocabulary. The number of times a joint descriptor was statistically significant in all the equations that constitute the GOM was counted in order to select the most important joint angles for each gesture vocabulary.

In order to evaluate the selection of the most meaningful joint angles, different combinations from the selected joint angles were used to train Hidden Markov Models (HMM) for gesture recognition using an "all-shots" approach. The motion data of sensors that provided at least one of the top three joint angles contributing the most in the response for the spine, arms, and legs parts motion was used for gesture recognition. Since one sensor provided three angles of one joint, all the joint angles of the sensor were used for training. The first combination to test for gesture recognition consisted of a minimal sensor configuration: the best sensor to measure the spine, another which was the best to measure the arms, and a third for the legs. If the recognition performance was low, an extra sensor was added to the configuration to improve the performance, or it was replaced by another of the top three sensors selected to measure its corresponding body location (spine, arms, or legs). The configuration that achieved the best performance was compared with the recognition performance obtained using all the joint angles of the sensors. The recognition performance, by using only a minimal set of two sensors, was also computed for comparison. This minimal set consisted of two hand-picked sensors, which provided the Euler joint angles of the right forearm (RFA) and hips (H). The sensor placed on the right forearm was chosen since most of the subjects in all datasets were right-handed, and the hips sensor was chosen because the origin of all movement of the spine starts from the hips.

To determine the best HMM setting for each gesture vocabulary, both ergodic and left-right topologies were tested, in addition to a different number of hidden states. The performance metric used consisted of the F-score. In the training phase of HMM, each professional gesture $\mathrm{G}_{v,c}$, where $v \in [1,4]$ indicates the gesture vocabulary and $c \in \mathbb{N}$ the gesture of the G_v, is associated to an HMM. The set of models for all gestures for every gesture vocabulary is $G_{v\in[1,4]} = \{HMM_c\}_{c\in\mathbb{N}}$.

3.3.2. Prediction of Joint Angles Trajectories

For evaluating the forecasting performance of the GOM models, the joint angle sequences of each gesture were simulated by solving the simultaneous equation system of the GOM. The models forecasted one time frame per iteration, then, after forecasting all the time frames of the gesture, the simulated gesture was compared with the original for evaluation. Consequently, their forecasting ability was evaluated by computing Theil's inequality coefficient (U) along with its decompositions: bias proportion ($\mathrm{U_B}$), variance proportion ($\mathrm{U_V}$), and covariance proportion ($\mathrm{U_C}$).

A sensitivity analysis was conducted to investigate the reaction of the models after a shock occurred in one of their variables. For this analysis, a disturbance of 80% was applied only in the first two frames of the gesture, then the whole gesture was forecasted. This analysis aimed to simulate the situation where subjects were exposed to external forces that affected their performance or made the workers assume awkward postures that increased the risk of injury.

4. Experimental Results

4.1. Statistical Significance of Motion Descriptors

Here, an example of a joint angle motion equation for one gesture from each vocabulary will be provided. These examples are offered to enable visualization of the coefficients and p-values of the different assumptions that compose the equation, where some variables need to remain dynamic and others static. The first example is for the equation of the gesture $\mathrm{G}_{1,1}$ (grab an electronic card from a container) for the joint angle RA_Y, which is the joint angle of the right arm on the Y-axis:

$$\begin{aligned} RA_Y(t) = {} & \underbrace{(-86.76)LSH1_Y(t-1)}_{p=0.01} + \underbrace{(-169.03)LSH1_Z(t-1)}_{p=0.001} + \\ & \underbrace{(88.48)LSH2_X(t-1)}_{p=0.008} + \underbrace{(-67.38)RSH1_Y(t-1)}_{p=0.001} + \\ & \underbrace{(-142.13)RSH1_Z(t-1)}_{p=0.002} + \cdots + \underbrace{(-2.18)RA_X(t-1)}_{p=0.508} \end{aligned} \tag{13}$$

By doing a statistical analysis of Equation (13), the p-values show intra-limb serial mediations with the joint angles on the Y and Z-axis of the left shoulder ($LSH1$) and intra-limb non-serial mediation with the right shoulder ($RSH1$). In the last equation, it should be noted that there is no intra-joint association shown by the p-value of the RA_X, and although it is not illustrated in the equation, there is no inter-limb synergy either. These results make sense since most of this motion is highly dependent on movements of the shoulders. Consequently, it is the reason that shoulders are statistically significant for the equation of RA_Y. The second example is the equation for $G_{2,3}$ (Hold the bucking bar) for the joint angle of the neck on the X-axis (N_X):

$$\begin{aligned} N_X(t) = \underbrace{(-1.2)N_Y(t-1)}_{p=0.001} + \underbrace{(-0.47)N_Z(t-1)}_{p=0.001} + \\ \underbrace{(-0.01)S2_X(t-1)}_{p=0.002} + \underbrace{(-0.02)S2_Y(t-1)}_{p=0.001} + \\ \underbrace{(-0.01)S3_X(t-1)}_{p=0.001} + \cdots + \underbrace{(0.01)H_X(t-1)}_{p=0.84} \end{aligned} \tag{14}$$

Equation (14) indicates that there is an intra-joint association with N_Y and N_Z, and an intra-limb serial mediation with the $S3$. There is an intra-limb non-serial mediation with $S2$, but not with H. For the gesture of holding a bucking bar to counteract a rivet, it is necessary to bend forward on the X-axis and Y-axis, which corresponds to what Equation (14) shows, that is to say, that joint angles from $S2$ and $S3$ on the X and Y-axis are statistically significant and contribute to gesture. Moreover, for this gesture, the subject needed to rotate the neck to see where to place the bucking bar; therefore, this matches with the intra-joint association indicated by the p-value of N_Y and N_Z.

The next equation is an example of gesture $G_{3,2}$ (shape the carafe curves) for the joint angle of the left shoulder on the X-axis, representing the motion of the left clavicle ($LSH2_X$):

$$\begin{aligned} LSH2_X(t) = \underbrace{(0.15)LSH2_Y(t-1)}_{p=0.003} + \underbrace{(0.17)LSH2_Z(t-1)}_{p=0.016} + \\ \underbrace{(-0.02)LA_Y(t-1)}_{p=0.001} + \underbrace{(-0.36)RSH2_X(t-1)}_{p=0.001} + \\ \underbrace{(-1.05)RSH2_Z(t-1)}_{p=0.001} + \cdots + \underbrace{(-0.01)LFA_X(t-1)}_{p=0.731} \end{aligned} \tag{15}$$

Statistical analysis of the Equation (15) indicates an intra-joint association, intra-limb serial mediation with the left arm, and an inter-limb synergy with the right shoulder. In this gesture, both arms must cooperate to shape the carafe correctly. The joints angles from the right shoulder contribute to the response of the left shoulder, since with the right arm the glassblower shaped the curves of the carafe, while the left arm slowly rolled the blowpipe. The Equation (16) presents a gesture from the G_4, where the subject bent forward more than 60° for the joint angle $S3$ on the Y-axis ($S3_Y$):

$$\begin{aligned} S3_Y(t) = \underbrace{(2.13)S3_X(t-1)}_{p=0.007} + \underbrace{(-0.17)S3_Z(t-1)}_{p=0.001} + \\ \underbrace{(-0.91)H_X(t-1)}_{p=0.012} + \underbrace{(0.42)S1_Y(t-1)}_{p=0.001} + \\ \underbrace{(-3.24)S2_X(t-1)}_{p=0.001} + \cdots + \underbrace{(-0.06)HE_X(t-1)}_{p=0.061} \end{aligned} \tag{16}$$

The p-values show that there is a dependency on the intra-joint association assumption. The joint angles on the X-axis from the sensors $S3$, H, and $S2$ are statistically significant and have the highest coefficient values, which is to be expected since the spine moves on the X-axis in order to bend forward. Moreover, there is an intra-limb serial and non-serial mediation with joint angles on the Y-axis, except for H_Y.

The top ten variables that contributed the most in the gestures of each gesture vocabulary are illustrated in Tables 1–4. From these joint angles, as mentioned in the methodology, different sets are used for gesture recognition. The results are shown in Section 4.2.

Table 1. G_1: Televison assembly.

***p*-Value < 0.05**					
Spine		**Arms**		**Legs**	
Variable	**Count**	**Variable**	**Count**	**Variable**	**Count**
$S1_Z$	49	LA_X	56	RUL_Y	32
$S2_Z$	47	$RSH1_X$	55	LUL_Z	32
H_Y	46	$RSH2_Y$	55	LUL_Y	31
H_Z	45	$RSH1_Z$	54	RL_Y	31
N_Y	44	$RSH2_Z$	53	LUL_X	29
$S1_X$	43	$RSH2_X$	53	LL_X	29
H_Z	42	RA_Y	49	RL_X	29
N_X	42	LFA_Z	48	H_X	29
$S1_Y$	41	$LSH1_X$	46	RUL_X	29
$S3_X$	41	LFA_X	42	RUL_Z	29

Table 2. G_2: Airplane assembly.

***p*-Value < 0.05**					
Spine		**Arms**		**Legs**	
Variable	**Count**	**Variable**	**Count**	**Variable**	**Count**
$S3_X$	209	$LSH2_X$	243	LUL_Z	39
$S3_Y$	205	$LSH1_X$	236	RUL_X	39
$S2_X$	202	RA_Z	230	H_X	38
H_Z	202	LFA_X	229	LL_X	38
H_X	201	RFA_Y	227	LUL_Y	38
S_X	201	LA_Y	224	LL_Y	37
$S1_Y$	197	LA_Z	217	LL_Z	37
$S1_Z$	193	$RSH1_X$	217	RL_X	36
$S3_Z$	193	LFA_Y	216	RL_Y	36
N_Y	193	LFA_Z	212	RL_Z	36

Table 3. G_3: Glassblowing.

***p*-Value < 0.05**					
Spine		**Arms**		**Legs**	
Variable	**Count**	**Variable**	**Count**	**Variable**	**Count**
$S3_X$	155	$LSH2_Y$	99	H_Y	65
$S3_Y$	155	RA_X	92	LL_Z	63
$S3_Z$	149	RFA_Z	90	LL_Y	62
$S2_X$	118	$LSH2_X$	89	RL_X	60
$S2_Z$	116	$RSH1_Z$	88	RL_Y	60
$S2_Y$	110	$LSH1_Z$	86	H_Z	59
$S1_Y$	105	$RSH1_Y$	85	LL_X	59
$S1_X$	102	$RSH2_X$	85	RUL_Y	58
$S1_Z$	93	LA_Y	84	RUL_Z	58
N_X	89	$LSH2_Z$	84	LUL_Y	57

Table 4. G_4: Motion primitives based on EAWS.

p-Value < 0.05					
Spine		**Arms**		**Legs**	
Variable	**Count**	**Variable**	**Count**	**Variable**	**Count**
$S3_Z$	332	$LSH1_X$	534	RUL_Z	474
$S2_Y$	330	LA_X	533	RUL_Y	473
$S2_Z$	330	$RSH1_X$	523	LUL_Y	472
$S3_X$	326	$LSH1_Y$	520	RL_X	468
$S3_Y$	316	LFA_X	520	LL_X	465
$S2_X$	311	$RSH2_X$	518	LUL_X	461
HE_Z	279	$RSH1_Y$	516	LUL_Z	457
S_Z	264	RA_X	514	RUL_X	456
H_Y	261	$RSH1_Z$	508	LFT_Z	455
N_Z	258	LA_Y	507	RFT_Y	455

4.2. Validation of the Joint's Selection

In this section, the recognition performances achieved by the different sets of sensors are reviewed and compared. Table 5 summarizes the results obtained when using different sensors and shows the configuration selected according to the GOM (SS), which achieved the best recognition performance. The gesture vocabulary G_1, as mentioned earlier, was composed of three gestures, each with 106 repetitions. HMMs with seven hidden states achieved the best recognition performance, trained with the joint angles provided by the sensors S1, LA, and RUL, selected by using the GOM. The precision and recall achieved with each configuration of sensors to recognize gestures from G_1 are illustrated in Tables 6 and 7.

Table 5. Recognition performance with each configuration of sensors. MS: Minimal set of two sensors. SS: Selected sensors by using the GOM.

Gesture Vocabularies	*N*° Classes	Sensors	F-Score (%)
G_1: TV assembly	3	MS: H and RFA	95.59
		SS: S1, LA, RUL	96.84
		All sensors	93.39
G_2: Airplane assembly	3	MS: H and RFA	88.89
		SS: S3, S2, LSH1, LSH2, RA, LUL, RUL	94.33
		All sensors	72.02
G_3: Glassblowing	5	MS: H and RFA	88.03
		SS: S3, LSH2, H, RFA	94.70
		All sensors	80.68
G_4: Motions based on EAWS	28	MS: H and RFA	73.85
		SS: S2, LA, RSH1, RUL, LFA	91.77
		All sensors	84.88

Table 6. Recall achieved for G_1 using HMMs.

Sensors	**Recall (%)**		
	$G_{1,1}$	$G_{1,2}$	$G_{1,3}$
MS: H and RFA	97.17	95.28	94.34
SS: S1, LA, RUL	94.34	99.06	97.17
All sensors	92.45	95.28	92.45

Table 7. Precision achieved for G_1 using HMMs.

Sensors	Precision (%)		
	$G_{1,1}$	$G_{1,2}$	$G_{1,3}$
MS: H and RFA	95.37	96.19	95.24
SS: S1, LA, RUL	98.04	95.45	97.17
All sensors	94.23	91.82	94.23

G_2 contained three gestures with 10 to 12 repetitions each. HMM with eight hidden states achieved the best performance. The SS sensor configuration had the best F-score, as shown in Table 5, 5.44% more than the set with the two sensors, and 22.31% more than the set with all the sensors. Tables 8 and 9 shows the recognition performance for G_2 with each sensor configuration, where the best precision and recall was achieved by the set SS.

Table 8. Recall achieved for G_2 using HMMs.

Sensors	Recall (%)		
	$G_{2,1}$	$G_{2,2}$	$G_{2,3}$
MS: H and RFA	83.33	83.33	100.00
SS: S3, S2, LSH1, LSH2, RA, LUL, RUL	100.00	83.33	100.00
All sensors	66.67	50.00	100.00

Table 9. Precision achieved for G_2 using HMMs.

Sensors	Precision (%)		
	$G_{2,1}$	$G_{2,2}$	$G_{2,3}$
MS: H and RFA	83.33	83.33	100.00
SS: S3, S2, LSH1, LSH2, RA, LUL, RUL	85.71	100.00	100.00
All sensors	57.14	60.00	100.00

G_3 consisted of five gestures with 10 to 35 repetitions for each. For this gesture vocabulary, HMM with four states modeled the best gestures and yielded the best recognition performance. This performance is illustrated in Tables 10 and 11. The configuration of sensors selected using GOM improved the overall F-score by at least 6% over the other sets. The G_4 was composed of the 28 motion primitives based on EAWS, where each exposed the subjects to different ergonomics risks concerning the posture. There are 30 repetitions for each motion, and HMM with seven states modeled the best gestures from G_4. For the gesture recognition of the 28 classes, the set SS yielded the higher F-score (91.77%), average precision (91.90%) and recall (92.33%), over the minimized set of two sensors and the set with all sensors. The minimized set achieved an average precision of 74.16% and an average recall of 77.31%. By using all the sensors for the recognition, an average precision of 84.76% and an average recall of 86.46% was achieved.

Table 10. Recall achieved for G_3 using HMMs.

Sensors	Recall (%)				
	$G_{3,1}$	$G_{3,2}$	$G_{3,3}$	$G_{3,4}$	$G_{3,5}$
MS: H and RFA	100.00	100.00	72.72	70.00	94.29
SS: S3, LSH2, H, RFA	83.33	95.45	100.00	100.00	97.14
All sensors	70.00	100.00	45.45	80.00	97.14

Table 11. Precision achieved for G_3 using HMMs.

Sensors	Precision (%)				
	$G_{3,1}$	$G_{3,2}$	$G_{3,3}$	$G_{3,4}$	$G_{3,5}$
MS: H and RFA	90.90	95.65	100.00	70.00	91.67
SS: S3, LSH2, H, RFA	100.00	95.45	81.82	100.00	97.14
All sensors	77.78	95.65	100.00	80.00	82.93

Performance Analysis of Selected Sensors Sets

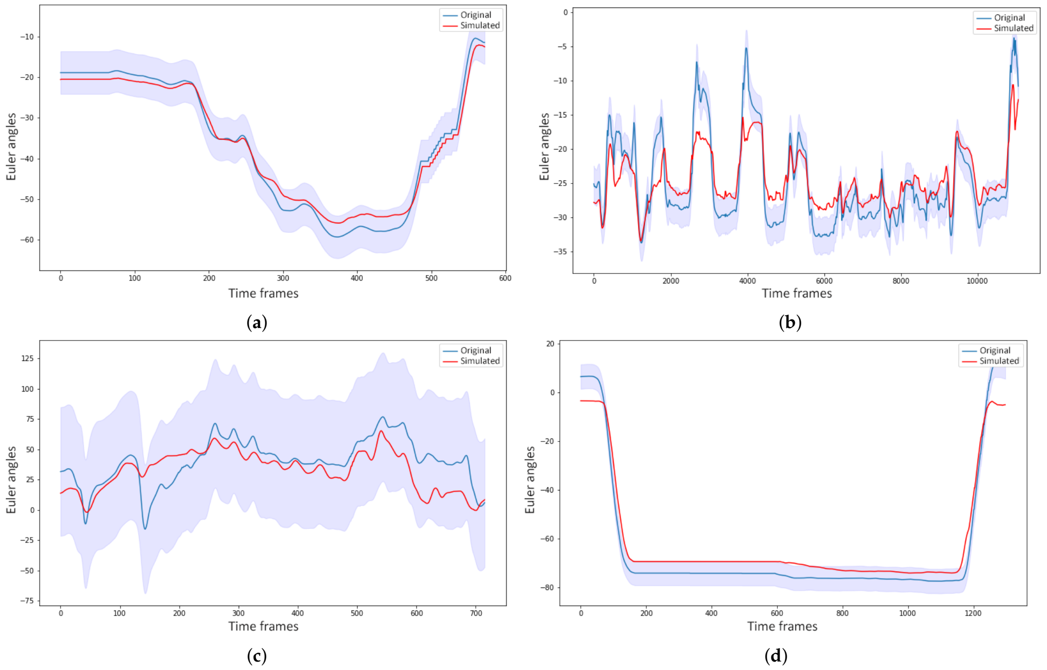
The relevance of the sensors selected for G_1 in the execution of the three gestures was proven due to the high recognition performance achieved. By observing the results for G_1 in Tables 6 and 7, it became apparent that the three sensors chosen improved the precision of the recognition and the recall of the gestures $G_{1,2}$ and $G_{1,3}$. In the case of $G_{1,1}$, the two sensors configuration had the best recall. Overall, the selected sensors had the best performance, with at least +1.2%. From the four gesture vocabularies, G_1 had the best performance for gesture recognition, which could be due to the low inter-class variance between the three gestures.

G_2 was the gesture vocabulary with the highest number of sensors selected. The reason could be because the gestures in this vocabulary were more complex and more prolonged. The most complicated gesture to model and recognize was $G_{2,2}$, which was expected since it is the gesture that could vary the most in its execution (high intraclass variance) from among the three gestures. The subject did not prepare the material in exactly the same way for each iteration. The subject was slower in some iterations than others since he required more time to adjust the pneumatic hammer or needed to prepare more rivets. The low intra-class variance could be because of the way the gestures were executed, which depended on the locations where the worker was going to fasten the metal plate with the rivets. For the recording of G_2, there was only one airplane structure to build, and there were no iterations where the subject placed the pneumatic hammer in the same location more than once.

The sensors selected for recognition of gestures from G_3 were validated by achieving a high recognition performance of the five gestures. By analyzing the results in Tables 10 and 11, the recall is improved in most gestures using the set SS, since the selected sensors capture the motion better. Regarding precision, the set SS improved it for $G_{3,4}$ and $G_{3,5}$, but then it decreased in comparison with the minimized set for $G_{3,1}$. This could be because the information provided by the sensors S3 and LSH2 generated similar patterns between the gestures $G_{3,1}$ and $G_{3,3}$. The minimized set had the worse precision and recall for $G_{3,4}$. The reason could be because of the lack of information on the motion of the shoulders. According to GOM, the shoulders contribute most to executing this gesture. Four out of the five gestures in this vocabulary generated similar patterns on the shoulder and arms. Still, there was low intra-class variation because of the high level of the subject's dexterity, as an expert in glassblowing. In addition, the subject used a metal structure for shaping the carafe that limited any potential freedom in the gesture performances. Finally, for G_4, a maximum F-score of 91.77% was recorded for the recognition of 28 motion primitives, using the selected sensors S2, LA, RS1, RUL and LFA. The poor performance of the minimized set was due to its failure to discriminate between motions that vary only with regard to posture of the legs, while the poor performance using all the sensors can be explained by the multiple dimensions of the data.

4.3. Simulated Movements and Sensitivity Analysis

This section presents the results of the trajectory prediction and sensitivity analysis. Figure 8 illustrates one example of a simulated gesture and the original from each gesture vocabulary, with confidence bounds of 95%. Figure 8b,c show that the models could capture the patterns generated on the motion of the spine by the task of buckling a rivet and the motion of the forearm while the glassblower was rotating the blowpipe. For more quantitative measurement of the forecasting performance, the Theil inequality coefficient, its decompositions, and the root mean square error were computed. Table 12 shows the forecasting performance for one gesture of each vocabulary on three Euler angles of a joint used during the execution of the gesture. By observing the Figure 8 and Table 12 alone, it can be assumed that the forecasting performance was good for these gestures. The original and simulated values were close to each other, and the simulated values were mostly inside the confidence bounds.



(**a**) (**b**)

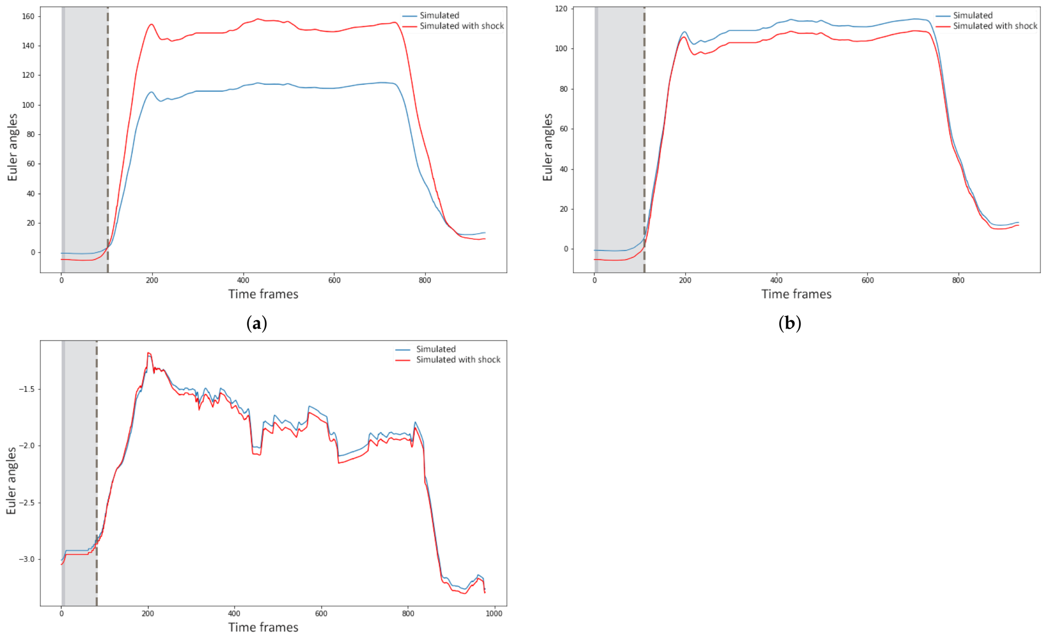
(**c**) (**d**)

Figure 8. Examples of simulated gestures, their original gesture, and confidence bounds of 95% (**a**) Simulation of the gesture $G_{1,3}$ on the joint angle LA_X; (**b**) The simulated joint angle sequence of S_Z for $G_{2,3}$; (**c**) Simulation of LFA_Y for the gesture $G_{3,1}$; (**d**) Forecasting of RA_X for $G_{4,9}$, which consists of raising the forearms above the shoulder level.

In Figure three examples of shocks applied to different variables for the sensitivity analysis are illustrated. Figure 9a,b show the forecasting behavior of the model of the joint angle LA_X for the gesture of raising the hands above the shoulder level. In Figure 9a a shock was applied on the joint angles of LSH2, and in Figure 9b, it was on the joint angles of RSH2. It is apparent that applying a shock to the left shoulder affected the motion of the left arm far more than applying it to the right shoulder, due to the strong mediation of the left shoulder over the left arm motion. Figure 9c shows the simulated motion of $S2_Y$ when the subjects rotated their torso to the left. The shock in this case was applied to the joint angles of H. It can be seen from the figure that the model was able to adapt fast and, indeed, in less than 1 s (90 frames), which indicated low sensitivity of the model to external disturbance. However, there was still a small variation in the forecasting if compared to the simulated gesture forecast without shocks.

Table 12. Forecasting performance of one gesture for each gesture vocabulary.

Gestures	Joint Angles	Theil Inequality U	Bias Proportion U_B	Variance Proportion U_V	Covariance Proportion U_C	RMSE
$G_{1,3}$	$LSH1_X$	0.0174	0.2499	0.0030	0.7483	0.0958
	$LSH1_Y$	0.0069	0.0000	0.0021	0.9996	0.0078
	$LSH1_Z$	0.0147	0.0006	0.0001	1.0010	0.0083
$G_{2,1}$	$RSH2_X$	0.0939	0.0002	0.0769	0.9230	0.0648
	$RSH2_Y$	0.0142	0.0000	0.0000	1.0001	0.0075
	$RSH2_Z$	0.0247	0.0002	0.0018	0.9981	0.0093
$G_{3,4}$	$LSH2_X$	0.2061	0.2786	0.0275	0.6947	0.2139
	$LSH2_Y$	0.3958	0.2327	0.0038	0.7645	0.1821
	$LSH2_Z$	0.3662	0.4919	0.1726	0.3361	0.6323
$G_{4,3}$	$S2_X$	0.0077	0.0290	0.0187	0.9531	0.0742
	$S2_Y$	0.0351	0.0906	0.2100	0.7001	0.1434
	$S2_Z$	0.0115	0.0692	0.0599	0.8717	0.0776



(**a**)

(**b**)

(**c**)

Figure 9. Simulated joint angles with and without disturbance of 80% on the two initial time frames. (**a**) Simulation of the joint angle LA_X with a disturbance on the joint angles of LSH2 (blue) and without (red); (**b**) Simulated joint angle sequence of LA_X with a disturbance on the joint angles of RSH2 (blue) and without (red); (**c**) Simulation of the joint angle $S2_Y$ with a disturbance on the joint angles of H (blue) and without (red).

5. Discussion

This paper evaluates GOM's feasibility to model industrial workers and subjects' dynamics and select the joint angles that best represent the gesture vocabulary, and predict their joint angles' trajectory. The statistical analysis conducted on the GOM models permitted identification of the joint angles that contributed most to the execution of the gestures of each vocabulary. For validation, these joint angles were then used for gesture recognition. These results demonstrate the potential of the selected set of sensors for a posture-based ergonomic analysis. By only using the data of the selected sensors, it was possible to discriminate accurately between different professional gestures and motion primitives where various postures of the spine, arms, and legs were assumed. The recognition of these changes in posture are clearly useful for ergonomic analysis of professional gestures. By applying a whole gesture to the trained HMMs from this vocabulary, the models could tell whether an awkward motion primitive is performed and which body part causes this ergonomic risk (i.e., spine, legs, or arms).

By solving the simultaneous equations that compose the GOM, it was possible to accurately forecast the modeled gesture, using Euler joint angles as input. Moreover, the models are tolerant to small variations in the gestures and offsets between same class gestures, which could be due to different recording conditions (different subjects or different recording days). Regarding the sensitivity analysis, the models showed low sensitivity to external disturbances, with only a small variation in the forecasting from that of a simulation without shocks. The response of the models to the shocks applied on different variables could be useful for detecting any physical strain (e.g., on the shoulders or lower back) or a load that affects the workers' performance and increases the ergonomic risk of the motion.

The industry has used ergonomic evaluations based on joint angle thresholds widely, due to their practicality. Previous studies have applied these evaluations in their analysis, where their only contribution was to fill them automatically using motion capture technologies [10–13]. An ergonomic analysis using such an approach can result in oversimplicity and ignore other potential risks workers are exposed to (e.g., external forces and dangerous movements). Menytchas [17] expanded such ergonomic analysis by examining the kinematics and kinetics of professional movements to identify joints that accumulate the most strain. The kinetic descriptors used in that study, however, did not allow for accurate discrimination between dangerous motions with small variations in the posture; moreover, they do not analyze the dynamics of movements, unlike the present study, which allowed for a good recognition performance and distinction between motions of different ergonomic risk.

In this study, GOM was proved to be useful for ergonomic analysis of professional motions. In comparison with the approach taken in the previous study by Manitsaris [14], a more in-depth analysis was conducted over the dynamic relationships of body parts, including more assumptions in the mathematical representation of each body joint motion. This gave insight into the influence of all body parts that work together to execute a specific movement and into devising helpful strategies to address ergonomic hazards, such as optimizing workspaces. Moreover, the methodology which was followed allowed the selection of the most meaningful joint angles for gesture recognition, improving the recognition performance considerably.

Despite the good performances and contributions achieved, this study highlighted some limitations regarding the use of inertial sensors in real workplace scenarios. Inertial sensors can offer precise and reliable measurements to study human motion; however, the degree of this precision and reliability depends on the site, movements, and tools handled during the performance. For instance, in the recording for the gesture vocabularies G_2 and G_3, workers used plastic gloves or did not wear the gloves that come with the inertial suit in order to avoid disturbances in the measurements. For this study, the motion data needed to be pre-processed after the recording to remove disturbances and drifts that could affect the results of the method.

6. Conclusions

From the literature reviewed, most of the studies used inertial sensors for quantifying the intensity, repetition, and duration of extreme motions and postures. The ability to extract information about work content from kinematics data is underutilized. Industrial workers perform complex professional gestures that contain crucial information about ergonomic risks. In this paper, not only was the contribution of every body joint in the execution of a specific professional gesture statistically estimated, but how they all operationally cooperate was modeled using GOM, and, in addition, their motion trajectories were accurately predicted using the trained models. GOM is based on state-space representations and consists of a simultaneous equation system of differential equations for all body body parts. The most significant joint motions for each gesture vocabulary were selected based on their statistical significance in the GOM models. The selection was then validated by achieving a high recognition performance of gestural time-series, which was modeled using continuous HMM. Four datasets were created for this work that contain professional gestures recorded under real conditions in factories and in a laboratory environment. The forecasting performance of the models was evaluated by comparing the simulated gestures with their original values. According to the Theil inequality coefficient and its decompositions, the performance of the models can be considered accurate.

Analyzing the response of the models to external disturbances and identifying the body joints to enable tracking for ergonomic monitoring could be useful for faster and more efficient evaluation of workers' gestures. Furthermore, the models could be used for ergonomic risk prevention. They could detect patterns in the motion trajectories that imply exposure to an ergonomic risk factor (e.g., workers are bending their torso or raising their arms beyond a level that could be considered ergonomically safe).

Lastly, using a full-body mocap suit in an industrial context has several difficulties. This study contributes to the literature by identifying the minimum motion descriptors to measure. This allows for the use of less intrusive technologies, such as smartphones and smartwatches, to measure these same motion descriptors. Future work will consist of adding kinetic measures to the assumptions that GOM models are composed of (e.g., joint moments), to complement the kinematic information, and will consider the effect of loads on the kinetic variables, which could indicate worker exposure to higher ergonomic risk.

Author Contributions: Conceptualization, B.E.O.-P. and S.M.; methodology, B.E.O.-P.; validation, B.E.O.-P., S.M. and A.G.; formal analysis, B.E.O.-P. and D.M.; resources, S.M.; data curation, B.E.O.-P. and D.M.; writing—original draft preparation, B.E.O.-P.; writing—review and editing, B.E.O.-P., S.M., D.M. and A.G.; funding acquisition, S.M. and A.G. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results has received funding by the EU Horizon 2020 Research and Innovation Programme under Grant Agreement No. 820767, CoLLaboratE project, and Grant No. 822336, Mingei project.

Institutional Review Board Statement: The research related to human recordings complies with all the relevant national regulations and institutional policies, and has been approved by the authors' institutional review board or equivalent committee.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All recorded data are available upon request.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

WMSD	Work-related musculoskeletal disorders
GOM	Gesture Operational Model
IMUs	Inertial Measurement Units
HMMs	Hidden Markov Models
MS	Minimum set of two sensors
SS	Selected sensors by using a GOM

Appendix A. Description of the Datasets' Gestures

This appendix presents a detailed description of the gestures from the four gesture vocabularies recorded. The first gesture vocabulary (G_1) consists of three gestures executed by two workers from the TV assembly sector. They grabbed an electronic card from a container, took a wire from another, connected them, and placed them on a TV chassis. The first gesture is grabbing the electronic card from a container ($G_{1,1}$), the second consists of taking a wire from a second container ($G_{1,2}$), and the third one involves connecting the electronic card and wire and placing them on the TV chassis ($G_{1,3}$).

The second gesture vocabulary is composed of three gestures performed in the airplane assembly sector (G_2). Two workers were recorded, one performing riveting with a pneumatic hammer and the other holding a bucking bar to counteract the incoming rivet. The first gesture in this second vocabulary is riveting with the pneumatic hammer ($G_{2,1}$), while the second is preparing the pneumatic hammer and grabbing rivets ($G_{2,2}$), and the third involves positioning the bucking bar to counteract the incoming rivet ($G_{2,3}$).

The third gesture vocabulary contains five gestures performed by a glassblower when creating a water carafe (G_3). In the first gesture, the glassblower with a blowpipe grabs melted glass from an oven while rotating the blowpipe ($G_{3,1}$). For the second gesture, the glassblower holds a specific paper with his right hand and shapes the carafe's curves while seated in a metallic structure ($G_{3,2}$). The third gesture involves blowing through the metal blowpipe that holds the glass for the carafe ($G_{3,3}$). In the fourth gesture, the glassblower is shaping the carafe's neck with pliers while standing ($G_{3,4}$), and the fifth gesture concerns heating the glass of the carafe in the oven while rotating the blowpipe ($G_{3,5}$).

The last gestural vocabulary is related to 28 motion primitives performed in a laboratory (G_4), recorded from ten subjects. Each gesture exposed the subjects to a different level of ergonomic risk. According to EAWS, the ergonomic risk level depends on the torso, legs, and arms posture. The gestures here varied in the posture of the spine, legs, and arms. For the torso posture, in some gestures, the subjects bent more than 60°, rotated the torso, laterally bent to the left, or rotated their torso while leaning forward. The legs posture changes depending on whether the subject executes the gesture standing, sitting, or kneeling. Regarding the arms posture, the changes consist of raising their arms above shoulder level or keeping them down and having the arms stretched or bent 90°.

References

1. Ranney, D.; Wells, R.; Moore, A. Upper limb musculoskeletal disorders in highly repetitive industries: Precise anatomical physical findings. *Ergonomics* **1995**, *38*, 1408–1423. [CrossRef] [PubMed]
2. De Kok, J.; Vroonhof, P.; Snijders, J.; Roullis, G.; Clarke, M.; Peereboom, K.; van Dorst, P.; Isusi, I. *Work-Related Musculoskeletal Disorders: Prevalence, Costs and Demographics in the EU*; Technical Report; European Agency for Safety and Health at Work: Bilbao, Spain, 2019; doi:10.2802/66947. [CrossRef]
3. Chiasson, M.E.; Imbeau, D.; Major, J.; Aubry, K.; Delisle, A. Influence of musculoskeletal pain on workers' ergonomic risk-factor assessments. *Appl. Ergon.* **2015**, *49*, 1–7. [CrossRef] [PubMed]
4. Lynn, M.; Corlett, N. RULA: A survey method for the investigation of work-related upper limb disorders. *Appl. Ergon.* **1993**, *24*, 91–99.
5. Schaub, K.; Caragnano, G.; Britzke, B.; Bruder, R. The European Assembly Worksheet. *Theor. Issues Ergon. Sci.* **2013**, *14*, 616–639. [CrossRef]
6. Karhu, O.; Kansi, P.; Kuorinka, I. Correcting working postures in industry: A practical method for analysis. *Appl. Ergon.* **1977**, *8*, 199–201. [CrossRef]
7. Pascual, S.A.; Naqvi, S. An investigation of ergonomics analysis tools used in industry in the identification of work-related musculoskeletal disorders. *Int. J. Occup. Saf. Ergon.* **2008**, *14*, 237–245. [CrossRef]
8. Snook, S.H.; Ciriello, V. The design of manual handling tasks: Revised tables of maximum acceptable weights and forces. *Ergonomics* **1991**, *34*, 1197–1213. [CrossRef]
9. David, G.C. Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders. *Occup. Med.* **2005**, *55*, 190–199. [CrossRef]
10. Busch, B.; Maeda, G.; Mollard, Y.; Demangeat, M.; Lopes, M. Postural optimization for an ergonomic human-robot interaction. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 2778–2785. [CrossRef]
11. Manghisi, V.M.; Uva, A.E.; Fiorentino, M.; Bevilacqua, V.; Trotta, G.F.; Monno, G. Real time RULA assessment using Kinect v2 sensor. *Appl. Ergon.* **2017**, *65*, 481–491. [CrossRef]
12. Plantard, P.; Shum, H.P.; Le Pierres, A.S.; Multon, F. Validation of an ergonomic assessment method using Kinect data in real workplace conditions. *Appl. Ergon.* **2016**, *65*, 562–569. [CrossRef]

13. Vignais, N.; Miezal, M.; Bleser, G.; Mura, K.; Gorecky, D.; Marin, F. Innovative system for real-time ergonomic feedback in industrial manufacturing. *Appl. Ergon.* **2013**, *44*, 566–574. [CrossRef] [PubMed]
14. Manitsaris, S.; Senteri, G.; Makrygiannis, D.; Glushkova, A. Human movement representation on multivariate time series for recognition of professional gestures and forecasting their trajectories. *Front. Robot. AI* **2020**, *7*, 1–20. [CrossRef] [PubMed]
15. Lu, T.W.; Chang, C.F. Biomechanics of human movement and its clinical applications. *Kaohsiung J. Med. Sci.* **2012**, *28*, S13–S25. [CrossRef]
16. Muller, A.; Pontonnier, C.; Robert-Lachaine, X.; Dumont, G.; Plamondon, A. Motion-based prediction of external forces and moments and back loading during manual material handling tasks. *Appl. Ergon.* **2020**, *82*, 102935. [CrossRef]
17. Menychtas, D.; Glushkova, A.; Manitsaris, S. Analyzing the kinematic and kinetic contributions of the human upper body's joints for ergonomics assessment. *J. Ambient Intell. Humaniz. Comput.* **2020**, *11*, 1–23. [CrossRef]
18. Faber, G.S.; Chang, C.C.; Kingma, I.; Dennerlein, J.T.; van Dieën, J.H. Estimating 3D L5/S1 moments and ground reaction forces during trunk bending using a full-body ambulatory inertial motion capture system. *J. Biomech.* **2016**, *49*, 904–912. [CrossRef]
19. Shojaei, I.; Vazirian, M.; Croft, E.; Nussbaum, M.A.; Bazrgari, B. Age related differences in mechanical demands imposed on the lower back by manual material handling tasks. *J. Biomech.* **2016**, *49*, 896–903. [CrossRef] [PubMed]
20. Wang, Z.; Mülling, K.; Deisenroth, M.P.; Ben Amor, H.; Vogt, D.; Schölkopf, B.; Peters, J. Probabilistic movement modeling for intention inference in human-robot interaction. *Int. J. Robot. Res.* **2013**, *32*, 841–858. [CrossRef]
21. Agarwal, A.; Triggs, B. Tracking articulated motion using a mixture of autoregressive models. In Proceedings of the Computer Vision—ECCV 2004, Prague, Czech Republic, 11–14 May 2004; pp. 54–65.
22. Devanne, M.; Berretti, S.; Pala, P.; Wannous, H.; Daoudi, M.; Del Bimbo, A. Motion segment decomposition of RGB-D sequences for human behavior understanding. *Pattern Recognit.* **2017**, *61*, 222–233. [CrossRef]
23. Lin, C.F.; Gross, M.; Ji, C.; Padua, D.; Weinhold, P.; Garrett, W.E.; Yu, B. A stochastic biomechanical model for risk and risk factors of non-contact anterior cruciate ligament injuries. *J. Biomech.* **2009**, *42*, 418–423. [CrossRef]
24. Donnell, D.M.S.; Seidelman, J.L.; Mendias, C.L.; Miller, B.S.; Carpenter, J.E.; Hughes, R.E. A stochastic structural reliability model explains rotator cuff repair retears. *Int. Biomech.* **2014**, *1*, 29–35. [CrossRef]
25. Fardi, B.; Schuenert, U.; Wanielik, G. Shape and motion-based pedestrian detection in infrared images: A multi sensor approach. In Proceedings of the IEEE Intelligent Vehicles Symposium, Las Vegas, NV, USA, 6–8 June 2005; Volume 2005, pp. 18–23. [CrossRef]
26. Binelli, E.; Broggi, A.; Fascioli, A.; Ghidoni, S.; Grisleri, P.; Graf, T.; Meinecke, M. A modular tracking system for far infrared pedestrian recognition. In Proceedings of the Intelligent Vehicles Symposium, Las Vegas, NV, USA, 6–8 June 2005; pp. 759–764. [CrossRef]
27. Schneider, N.; Gavrila, D.M. Pedestrian path prediction with recursive Bayesian filters: A comparative study. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; 8142 LNCS; Springer: Berlin/Heidelberg, Germany, 2013; pp. 174–183._18. [CrossRef]
28. Barth, A.; Franke, U. Where will the oncoming vehicle be the next second? In Proceedings of the 2008 IEEE Intelligent Vehicles Symposium, Eindhoven, The Netherlands, 4–6 June 2008; pp. 1068–1073. [CrossRef]
29. Zernetsch, S.; Kohnen, S.; Goldhammer, M.; Doll, K.; Sick, B. Trajectory prediction of cyclists using a physical model and an artificial neural network. In Proceedings of the 2016 IEEE Intelligent Vehicles Symposium (IV), Gothenburg, Sweden, 19–22 June 2016; pp. 833–838. [CrossRef]
30. Pool, E.A.; Kooij, J.F.; Gavrila, D.M. Using road topology to improve cyclist path prediction. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 289–296. [CrossRef]
31. Kooij, J.F.; Flohr, F.; Pool, E.A.; Gavrila, D.M. Context-based path prediction for targets with switching dynamics. *Int. J. Comput. Vis.* **2019**, *127*, 239–262. [CrossRef]
32. Quintero, R.; Parra, I.; Llorca, D.F.; Sotelo, M.A. Pedestrian path, pose and intention prediction through Gaussian process dynamical models and pedestrian activity recognition. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 1803–1814. [CrossRef]
33. Kucner, T.P.; Magnusson, M.; Schaffernicht, E.; Bennetts, V.H.; Lilienthal, A.J. Enabling flow awareness for mobile robots in partially observable environments. *IEEE Robot. Autom. Lett.* **2017**, *2*, 1093–1100. [CrossRef]
34. Sun, L.; Yan, Z.; Mellado, S.M.; Hanheide, M.; Duckett, T. 3DOF pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2017; pp. 5942–5948. [CrossRef]
35. Xue, H.; Huynh, D.Q.; Reynolds, M. SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1186–1194. doi:10.1109/WACV.2018.00135. [CrossRef]
36. Srikanth, S.; Ansari, J.A.; Ram, R.K.; Sharma, S.; Murthy, J.K.; Krishna, K.M. INFER: INtermediate representations for FuturE pRediction. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 4–8 November 2019; pp. 942–949. [CrossRef]
37. Rudenko, A.; Palmieri, L.; Herman, M.; Kitani, K.M.; Gavrila, D.M.; Arras, K.O. Human motion trajectory prediction: A survey. *Int. J. Robot. Res.* **2020**, *39*, 895–935. [CrossRef]

38. Best, G.; Fitch, R. Bayesian intention inference for trajectory prediction with an unknown goal destination. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Hamburg, Germany, 28 September–3 October 2015; pp. 5817–5823. doi:10.1109/IROS.2015.7354203. [CrossRef]
39. Lee, W.; Seto, E.; Lin, K.Y.; Migliaccio, G.C. An evaluation of wearable sensors and their placements for analyzing construction worker's trunk posture in laboratory conditions. *Appl. Ergon.* **2017**, *65*, 424–436. [CrossRef]
40. Peppoloni, L.; Filippeschi, A.; Ruffaldi, E.; Avizzano, C.A. A novel wearable system for the online assessment of risk for biomechanical load in repetitive efforts. *Int. J. Ind. Ergon.* **2014**, *52*, 1–11. [CrossRef]
41. Ryu, J.; Seo, J.; Jebelli, H.; Lee, S. Automated action recognition using an accelerometer-embedded wristband-type activity tracker. *J. Constr. Eng. Manag.* **2019**, *145*. [CrossRef]
42. Slaton, T.; Hernandez, C.; Akhavian, R. Construction activity recognition with convolutional recurrent networks. *Autom. Constr.* **2020**, *113*, 103138. [CrossRef]
43. Parsa, B.; Samani, E.U.; Hendrix, R.; Devine, C.; Singh, S.M.; Devasia, S.; Banerjee, A.G. Toward ergonomic risk prediction via segmentation of indoor object manipulation actions using spatiotemporal convolutional networks. *IEEE Robot. Autom. Lett.* **2019**, *4*, 3153–3160. [CrossRef]
44. Caramiaux, B.; Montecchio, N.; Tanaka, A.; Bevilacqua, F. Adaptive gesture recognition with variation estimation for interactive systems. *ACM Trans. Interact. Intell. Syst.* **2015**, *4*. [CrossRef]
45. Pavlovic, V.; Rehg, J.M.; MacCormick, J. Learning switching linear models of human motion. In Proceedings of the 13th International Conference on Neural Information Processing Systems, ACM, Hong Kong, China, 3–6 October 2001; pp. 942–948. [CrossRef]
46. Aksan, E.; Kaufmann, M.; Hilliges, O. Structured Prediction Helps 3D Human Motion Modelling. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 7144–7153. [CrossRef]
47. Pavllo, D.; Feichtenhofer, C.; Auli, M.; Grangier, D. Modeling Human Motion with Quaternion-Based Neural Networks. *Int. J. Comput. Vis.* **2020**, *128*, 855–872. [CrossRef]
48. Wang, J.; Tang, S. Time series classification based on arima and adaboost. In Proceedings of the International Conference on Computer Science Communication and Network Security (CSCNS2019), Sanya, China, 22–23 December 2020; Volume 309, p. 7. [CrossRef]
49. Bobick, A.F.; Wilson, A.D. A state-based approach to the representation and recognition of gesture. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 1325–1337. [CrossRef]
50. Holmes, E.E. Kalman filtering for maximum likelihood estimation given corrupted observations. *Natl. Mar. Fish. Serv.* **2003**, *22*, 929–932. [CrossRef]

frontiers
in Robotics and AI

HYPOTHESIS AND THEORY
published: 13 August 2020
doi: 10.3389/frobt.2020.00080

Human Movement Representation on Multivariate Time Series for Recognition of Professional Gestures and Forecasting Their Trajectories

Sotiris Manitsaris, Gavriela Senteri, Dimitrios Makrygiannis and Alina Glushkova*

Centre for Robotics, MINES ParisTech, PSL Université, Paris, France

Human-centered artificial intelligence is increasingly deployed in professional workplaces in Industry 4.0 to address various challenges related to the collaboration between the operators and the machines, the augmentation of their capabilities, or the improvement of the quality of their work and life in general. Intelligent systems and autonomous machines need to continuously recognize and follow the professional actions and gestures of the operators in order to collaborate with them and anticipate their trajectories for avoiding potential collisions and accidents. Nevertheless, the recognition of patterns of professional gestures is a very challenging task for both research and the industry. There are various types of human movements that the intelligent systems need to perceive, for example, gestural commands to machines and professional actions with or without the use of tools. Moreover, the *inter*class and *intra*class spatiotemporal variances together with the very limited access to annotated human motion data constitute a major research challenge. In this paper, we introduce the Gesture Operational Model, which describes how gestures are performed based on assumptions that focus on the dynamic association of body entities, their synergies, and their serial and non-serial mediations, as well as their transitioning over time from one state to another. Then, the assumptions of the Gesture Operational Model are translated into a simultaneous equation system for each body entity through State-Space modeling. The coefficients of the equation are computed using the Maximum Likelihood Estimation method. The simulation of the model generates a confidence-bounding box for every entity that describes the tolerance of its spatial variance over time. The contribution of our approach is demonstrated for both recognizing gestures and forecasting human motion trajectories. In recognition, it is combined with continuous Hidden Markov Models to boost the recognition accuracy when the likelihoods are not confident. In forecasting, a motion trajectory can be estimated by taking as minimum input two observations only. The performance of the algorithm has been evaluated using four industrial datasets that contain gestures and actions from a TV assembly line, the glassblowing industry, the gestural commands to Automated Guided Vehicles as well as the Human–Robot Collaboration in the automotive assembly lines. The hybrid approach State-Space and HMMs outperforms standard continuous HMMs and a 3DCNN-based end-to-end deep architecture.

Keywords: state-space representation, differential equations, movement modeling, hidden Markov models, gesture recognition, forecasting, motion trajectory

Edited by:
Shalabh Gupta,
University of Connecticut,
United States

Reviewed by:
Patrick Henaff,
Université de Lorraine, France
Pedro Neto,
University of Coimbra, Portugal

****Correspondence:***
Sotiris Manitsaris
sotiris.manitsaris@mines-paristech.fr

Specialty section:
This article was submitted to Sensor Fusion and Machine Perception, a section of the journal Frontiers in Robotics and AI

Received: *20 December 2019*
Accepted: *08 May 2020*
Published: *13 August 2020*

Citation:
Manitsaris S, Senteri G, Makrygiannis D and Glushkova A (2020) Human Movement Representation on Multivariate Time Series for Recognition of Professional Gestures and Forecasting Their Trajectories. Front. Robot. AI 7:80. doi: 10.3389/frobt.2020.00080

INTRODUCTION

Human motion analysis and recognition are widely researched from various scientific domains including Human–Computer Interaction, Collaborative Robotics, and Autonomous Vehicles. Both the industry and science face significant challenges in capturing the human motion, developing models, and algorithms for efficiently recognizing it, as well as for improving the perception of the machines when collaborating with humans.

Nevertheless, in factories, "we always start with manual work," as explained by Mitsuri Kawai, Head of Manufacturing and Executive Vice-President of Toyota (Borl, 2018). Therefore, experts from both collaborative robotics and applied ergonomics are always involved when a new collaborative cell is being designed. Nowadays, despite the significant progress in training robots by demonstration, automatizing the human tasks in mixed workspaces still remains the goal. However, those workspaces are not necessarily collaborative. For example, in a smart workspace, a machine that perceives and anticipates gestures and actions of the operator would be able to adapt its own actions depending on those of the operator, thus giving him/her the possibility to obtain ergonomically "green postures." Furthermore, automated guided vehicles (AGVs) should also be able to detect the intentions of the operator with the aim of collaborating with them, avoiding accidents and understanding gestural commands. Finally, in Industry 4.0, an important number of Creative and Cultural Industries, for example, in luxury goods manufacturing, still base their know-how on manual dexterity, no matter whether the operator is in collaboration with a machine or not. Therefore, human movement representation and gesture recognition constitute a mean for identifying the industrial know-how and transmitting it to the next generation of the operators.

From a scientific point of view, major research challenges are faced by scientists, especially when dealing with professional environments in an industrial context. Initially, there is an extremely limited access to motion data from real-life configurations. This is mainly due to acceptability issues from the operators or to limitations imposed by laws and regulations that protect the access to/use of personal data, for example, the "General Data Protection Regulation" in the European Union. Therefore, most of existing datasets include only gestures from the everyday life. Furthermore, when creating custom datasets with professional motion data, many practical, and environmental issues might occur, for example, variation in luminosity, various workspace with different geometries, camera in motion to record a person moving in space, and low availability of real experts. Additionally, the community of actions and gesture recognition deals with challenges that are related to intraclass and *inter*class variations (Fu, 2016). Frequent are the cases where a professional task involves gestures that have very similar spatiotemporal characteristics (*low interclass variation*) together with very important differences in the way different humans perform the same gesture (*high intraclass variation*). Finally, when applied to accident prevention, a small delay in predicting the action might be crucial.

The work presented in this paper contributes to the aforementioned challenges, through the proposition of a Gesture Operational Model (GOM) that describes how the body parts cooperate, to perform a situated professional gesture. The model is built upon several assumptions that determine the dynamic relationship between the body entities within the execution of the human movement. The model is based on the State-Space (SS) representation, as a simultaneous equation system for all the body entities is generated, composed of a set of first-order differential equations. The coefficients of the equation system are estimated using the Maximum Likelihood Estimation (MLE) method, and its dynamic simulation generates a dynamic tolerance of the spatial variance of the movement over time. The scientific evidence of the GOM is evaluated through its ability to improve the recognition accuracy of gestural time series that are modeled using continuous Hidden Markov Models (HMMs). Moreover, the system is dynamically simulated through the solution of its equations. Its forecasting ability is evaluated by comparing the similarity between the real and simulated motion data using two real observations for initializing the models as well as by measuring the Theil inequality coefficient and its decompositions.

The performance of the algorithms that implement the GOM, the recognition of gestures, and the forecasting of the motion trajectories are evaluated by recording four industrial real-life datasets from a European house-holding manufacturer, a glassblowing workshop, an AGV manufacturer, and a scenario in automotive industry. More precisely, the first dataset contains motion data with gestures and actions from a TV assembly line, the second from the creation of glass water carafes, the third gestural commands to mobile robots, and the fourth from a scenario of Human–Robot Collaboration in the automotive industry. The motion data used in our experiments are 2D positions that are exported from computer vision and the application of a deep-learning-based pose estimation using the OpenPose framework (Cao et al., 2019).

State of the Art presents a state of the art on human motion modeling, representation, and recognition. In *Methodology*, the whole methodology analysis, modeling, and recognition are presented, whereas in *Evaluation*, the different approaches in the evaluation of the ability of the models to simulate a gesture and forecast its trajectories are analyzed. In the same section, the accuracy of the proposed method is also presented. Finally, in *Discussion* and *Conclusion and Future Work*, a discussion and the future work and perspectives of the proposed methodology are described.

STATE OF THE ART

In professional environments, a movement can be defined as the displacement of the human body in space, whereas gesture is a form of non-verbal communication for interacting with a machine or manipulating a tool or object. In industry, professional gestures define the routine of workers. The non-physical interaction of the workers with collaborative machines is relying on an external perception layer that gets input from the ambient environment using, among others, human motion sensors, to understand their movement and adapt their behavior according to the humans. Whether machine or deep learning can be used to create an AI-based perception layer for collaborative machines. Machine learning has demonstrated an important number of applications in action and gesture recognition by using whether a probabilistic modeling of the phenomenon, and optionally a representation of it, or a template matching that makes use of a temporal rescaling of the input signal according to the reference. Finally, architectures of deep neural networks have recently seen a considerable number of applications with a high accuracy and precision.

Movement Modeling and Representation

Each body articulation is strongly affected by the movement of other articulations. Observing a person running brings evidence on the existing interdependencies between different parts of human body that need to move cooperatively for a movement to be achieved. Duprey et al. (2017) attempted to study those relationships by exploring the upper body anatomy models available and describe their applicability using multi-body kinematic optimization, mostly for clinical, and ergonomic uses. Biomechanics has also actively contributed to the study of human movement modeling by using Newtonian methods and approaches, especially in sports and physical rehabilitation (Zatsiorsky, 2000). The representation of the human movement with physical or statistical models provides with a simplified mathematical formalization of the phenomenon and approximates reality, e.g. through simulation and forecasting. State-Space (SS) is a statistical modeling that allows to stochastically represent a human movement through a reasoning over time that makes use of internal states. An SS representation is a mathematical model of a physical system as a set of input, output, and state variables related by first-order differential equations. Kalman filtering is a method that estimates and determines values for the parameters of a model. To represent human movement, Zalmai et al. (2015) used linear SS models and provided an algorithm based on local likelihood for detecting and inferring gesture causing magnetic field variations. Lech and Kostek (2012) used Kalman filtering to achieve hand tracking and presented a system based on camera and multimedia projector enabling a user to control computer applications by dynamic hand gestures. Finally, Dimitropoulos et al. (2016) presented a methodology for the modeling and classification of multidimensional time series by exploiting the correlation between the different channels of data and the geometric properties of the space in which the parameters of the descriptor lie by using a linear dynamic system (LDS). Here, multidimensional evolving data were considered as a cloud of points (instead of a single point) on the Grassmann manifold, and codebook is created to represent each multidimensional signal as a histogram of Grassmannian points, which is not always the case for professional gestures.

Machine Learning for Gesture Recognition

Template-Based Machine Learning

Template-based machine learning has been widely used for gesture recognition in the context of continuous real-time human–machine interaction. Dynamic Time Warping (DTW) is a template-based method that has been widely used for measuring the similarity between motion data. DTW makes it possible to find the optimal global alignment between two sequences. Bevilacqua et al. (2005), Bevilacqua et al. (2007), and Bevilacqua et al. (2009) successively developed a system based on DTW, the Gesture Follower, for both continuous gesture recognition and following, between the template or reference gesture, and the input or performed one. A single example allows the training of the system (Bobick and Wilson, 1997). During the performance, a continuous estimation of parameters is calculated in real time, providing information for the temporal position of the performed gesture. Time alignment occurs between the template and the performed gesture, as well as an estimation of the time progression within the template in real time. Instead, Psarrou et al. (2002) used the Conditional Density Propagation algorithm to perform gesture recognition and made sure that they will not get probabilities for only one model per time stamp. The experiments resulted to a relatively good accuracy for the time period conducted.

Model-Based Machine Learning

One of the most popular methods of model-based machine learning that has been used to model and recognize movement patterns are HMMs. Pedersoli et al. adopted this method (Pedersoli et al., 2014) to recognize in real-time static hand poses and dynamic hand gestures of American Sign Language. Sideridis et al. (2019) created a gesture recognition system for everyday gestures recorded with inertia measurement units, based on Fast Nearest Neighbors, and Support Vector Machine methods, whereas Yang and Sarkar (2006) chose to use an extension of HMMs. Vaitkevičius et al. (2019) used also HMMs with the same purpose, gesture recognition, for the creation of virtual reality installations, as well as Williamson and Murray-Smith (2002), who used a combination of HMMs with a dynamic programming recognition algorithm, along with the granular synthesis method for gesture recognition with audio feedback. In a more industrial context, Yang et al. (2007) used gesture spotting with HMMs to achieve efficient Human–Robot collaboration where real-time gesture recognition was performed with extended HMM methods like Hierarchical HMMs (Li et al., 2017). HMMs seem to be a solid approach, allowing to achieve satisfying results of gesture modeling and recognition and are suitable for real-time applications.

The aforementioned methodologies and research approaches permit the identification of what/which gesture is performed by giving a probability, but not how expressively the gesture has been performed. Caramiaux et al. (2015) extended the research, by implementing a sequential Monte Carlo technique to deal with expressivity. The recognition system, named Gesture Variation Follower, is being adapted to gesture expressive variations in real time. Specifically, in the learning phase, only one example per gesture is required. Then, in the testing (recognition) phase, time alignment is computed continuously, and expressive variations (such as speed and size) are estimated between the template and the performed gesture (Caramiaux, 2015; Caramiaux et al., 2015).

The model-based and template-based methods present an interesting complementarity and their combination in most of cases, give the possibility to achieve satisfying gesture recognition accuracy. However, when the probability given per class presents a high level of uncertainty, these methods need to be completed with an extra layer of control that will permit to take a final, more robust, decision about the probability of an observation to belong to each class. One of the goals of this work is to focus on the use of the SS method for human movement representation and modeling and to use this representation as the extra control layer to improve gesture recognition results.

Deep Architectures for Action Recognition

Deep Learning (DL) is another approach with increasing scientific evidence in action and gesture classification. Mathe et al. (2018) presented results on hand gesture recognition with the use of a Convolutional Neural Network (CNN), which is trained on Discrete Fourier Transform images that resulted from raw sensor readings. In Oyedotun and Khashman (2017), the authors proposed an approach for the recognition of hand gestures from the American Sign Language using CNNs and auto-encoders. 3DCNNs are used in Molchanov et al. (2015) to detect hand movements of drivers, and in Camgoz et al. (2016) continuously recognize gesture classes from the Continuous Gesture Dataset (ConGD), which is the larger user-independent dataset. A two-stage approach is presented in Li et al. (2015), which combines feature learning from RGB-D using CNNs with Principal Component Analysis (PCA) for selecting the final features. Devineau et al. (2018) used a CNN model and tested its performance on classifying sequential humans' tasks using hand-skeletal data as input. Shahroudy et al. (2016), wanting to improve their action recognition results and decrease the dependency in factors like lightning, background, and color clothing, used a recurrent neural network to model the long-term correlation of the features for each body part. For the same reason, Yan et al. (2018) proposed a model of dynamic skeletons called spatial–temporal graph convolutional network (ST-GCN). This Neural Network (NN) learns automatically the spatial and temporal patterns from the given data, minimizing the computational cost, and increasing the generalization capability. In other cases, action recognition was achieved using either 3DCNNs (Tran et al., 2015) or two stream networks (Simonyan and Zisserman, 2014). CNNs are the NNs used in all cases above, as they are the main method used for image pattern recognition.

The particularity of DL methods is that they require a big amount of data in order to be trained. In some applications, having access to an important volume of data might not be possible for various reasons. One application with extremely limited amount of data is the recognition of situated professional actions and gestures performed in an industrial context, such as in manufacturing, assembling lines, and craftsmanship. Deep NNs are powerful methods for pattern recognition with great accuracy results, but they present some limitations for real-time applications, which are linked to the computational power that is required for both training and testing purposes. In this paper, given the fact that the available examples per gesture class are also limited, it is assumed and proved that stochastic model-based machine learning can give better results than DL.

The aim of this paper is to get advantage of existing knowledge in machine learning, and more precisely in the stochastic modeling for the recognition of gestures and the forecasting of their motion trajectories. To underline the advantages of the method proposed we also compare its performance with results obtained with a recent DL-based end-to-end architecture.

Our Previous Work

Manitsaris et al. (2015b) previously defined an operational model explaining how the body parts are related to each other, which was used for the extraction of confidence bounds over the time series of motion data. In Manitsaris et al. (2015b), as well as in Volioti et al. (2016), the operational model has been tested on Euler angles. In this work, the operational model is expanded to the full body and is tested with various datasets with position data from various real-life situations that have more classes and users than in previous work.

METHODOLOGY

Overview

The motion capturing of the operators in their workplace is a major task. A number of professional gestural vocabularies are created to build the methodology and evaluate its scientific evidence. Although the proposed methodology (**Figure 1**) is compatible with various types of motion data, we opted for RGB sensors and, in most of cases, with 2D positions to avoid any interference between the operator and his/her tools or materials. Thus, RGB images are recorded for every gesture of the vocabulary, segmented into gesture classes, annotated, and then introduced to an external framework for estimating the poses and extracting the skeleton of the operators.

As shown in **Figure 2**, the GOM is based on a number of assumptions that describe the way the different entities of the human body cooperate to efficiently perform the gesture. The assumptions of the model refer to various relationships between the entities, which are: the *intrajoint association*, the *interlimb synergies*, the *intralimb mediation*, and the *transitioning* over time. Following the theory of the SS modeling, the GOM is translated into a *simultaneous equation system* that is composed of two first-order differential equations for each component (e.g., dimension X, Y for 2D or X Y, Z for 3D) of each body entity.

During the training phase, the motion data of the training dataset are used to compute the coefficients of the equation system using the MLE method but also to execute a supervised learning of the continuous HMMs. Moreover, the motion data are used to solve the simultaneous equation system and simulate the whole gesture, thus generating values for the state variables. Once the solution of the system is completed for all the gestures of the vocabulary, the forecasting ability of every model is evaluated using the Theil coefficients as well as their performance in comparison with the motion data of other gestures.

During the testing phase, the HMMs output their likelihoods, which are multiplied by a confidence coefficient when their maximum likelihood is under a threshold. Finally, a motion trajectory can be dynamically or statically forecasted at any time

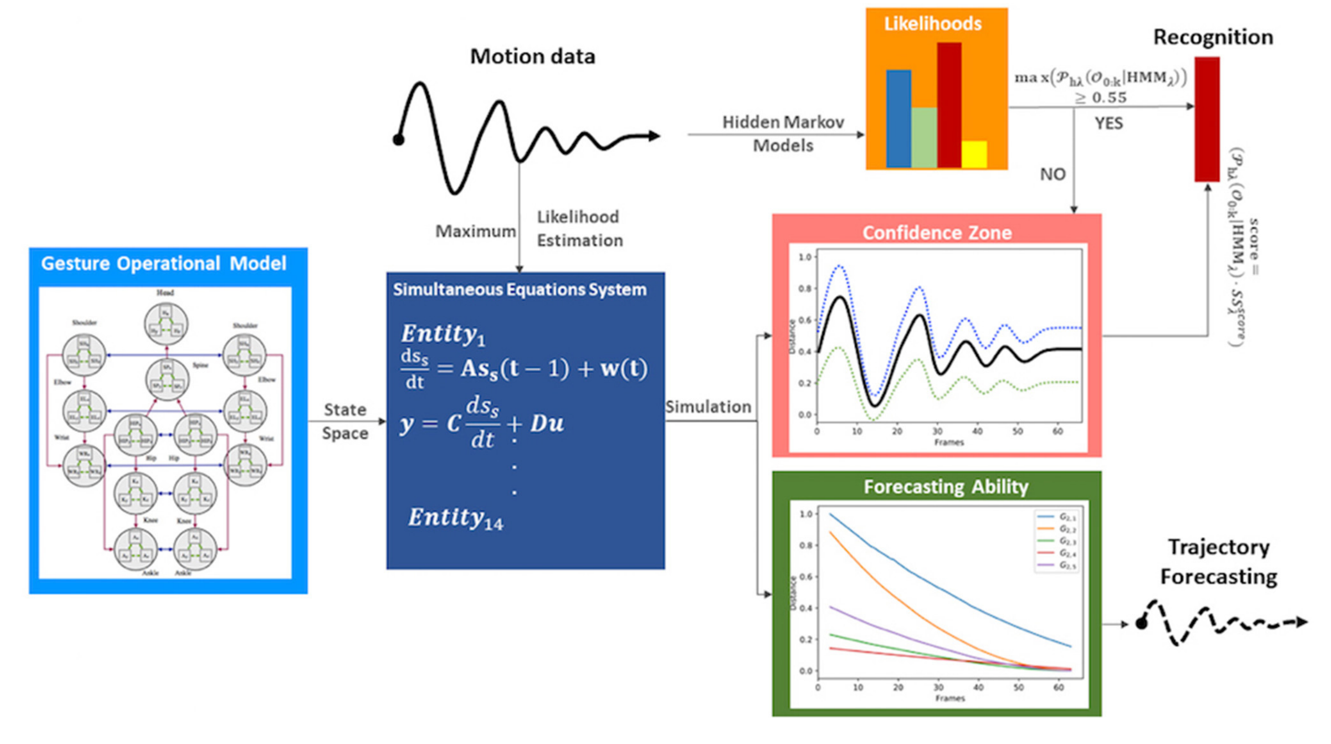

Motion data
Maximum Likelihood Estimation
Gesture Operational Model
State Space
Simultaneous Equations System
$Entity_1$
$\frac{ds_s}{dt} = As_s(t-1) + w(t)$
$y = C\frac{ds_s}{dt} + Du$
$Entity_{14}$
Hidden Markov Models
Likelihoods
$\max(\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_\lambda)) \geq 0.55$
YES
NO
Recognition
$score = (\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_\lambda) \cdot SS_\lambda^{score})$
Simulation
Confidence Zone
Forecasting Ability
Distance
Frames
Trajectory Forecasting

FIGURE 1 | Methodology pipeline.

by giving as input at least two time-stamp values from the real motion data.

Industrial Datasets and Gesture Vocabularies

The performance of the algorithms is evaluated by recording four industrial real-life datasets from a house-holding manufacturer, a glassblowing workshop, an AGV manufacturer, and an automotive industry. For each dataset, a gesture vocabulary has been defined in order to segment the whole procedure into small human motion units (**Table 2**).

The first gesture vocabulary (GV_1) includes four gestures where the operator takes the electronic card from one box and then takes a wire from another, connects them, and places them on the TV chassis. The gestures are performed in a predefined working space, in front of the conveyor and with the boxes placed on the left and right sides. However, the operator has a certain degree of variation in the way of executing the tasks, because the gestures are ample involving the whole body. Moreover, in order to avoid self-occlusions and scene occlusions, the camera is mounted on the top, which is not necessarily the optimal camera location for pose estimation algorithms, for example, OpenPose. Currently, in the factory, together with the operator who performs the actions of GV_1, there is also a second operator who will be progressively replaced by a collaborative robot.

The second gestural vocabulary proposes gestural commands for controlling an AGV. This dataset GV_2 contains five gestures involving mostly the arm and forearm. $G_{2,1}$ initiates the communication with the AGV, by shaking the palm, whereas $G_{2,2}$ and $G_{2,3}$ turn left and right the AGV by raising the respective arms. $G_{2,4}$ speeds up the AGV by raising three times the right hand, whereas $G_{2,5}$ speeds down the AGV by rolling the right hand away from the hips with a distance of around 20/30 cm. All gestures of GV_2 start and end with the *i*-pose.

The third gesture vocabulary GV_3 contains four gestures performed by a glassblower when creating a water carafe. The craftsman executes the gestures in a very limited space that is defined by a specific metallic construction. The craftsman puts the pipe on the metallic structure and performs various manipulations of the glass by using tools, such as pliers. Three out of four gestures are performed while the craftsman is sitting. More precisely, he starts by shaping the neck of the carafe with the use of pliers ($G_{3,1}$), then he tightens the neck to define the transition between the neck and the curved vessel ($G_{3,2}$), he holds in his/her right hand a specific paper and shapes the curves of the blown part ($G_{3,3}$), and finalizes the object and fixes the details by using a metallic stick ($G_{3,4}$). In general, the right hand is manipulating the tools while the left is holding and controlling the pipe. In parallel with $G_{3,2}$ and $G_{3,3}$, an assistant is helping and blowing promptly the pipe to permit the creation of the blown curved part.

The last dataset (GV_4) used in this paper is related to a real-life Human–Robot Collaboration scenario that has been recorded

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/output.mdwhwhoch

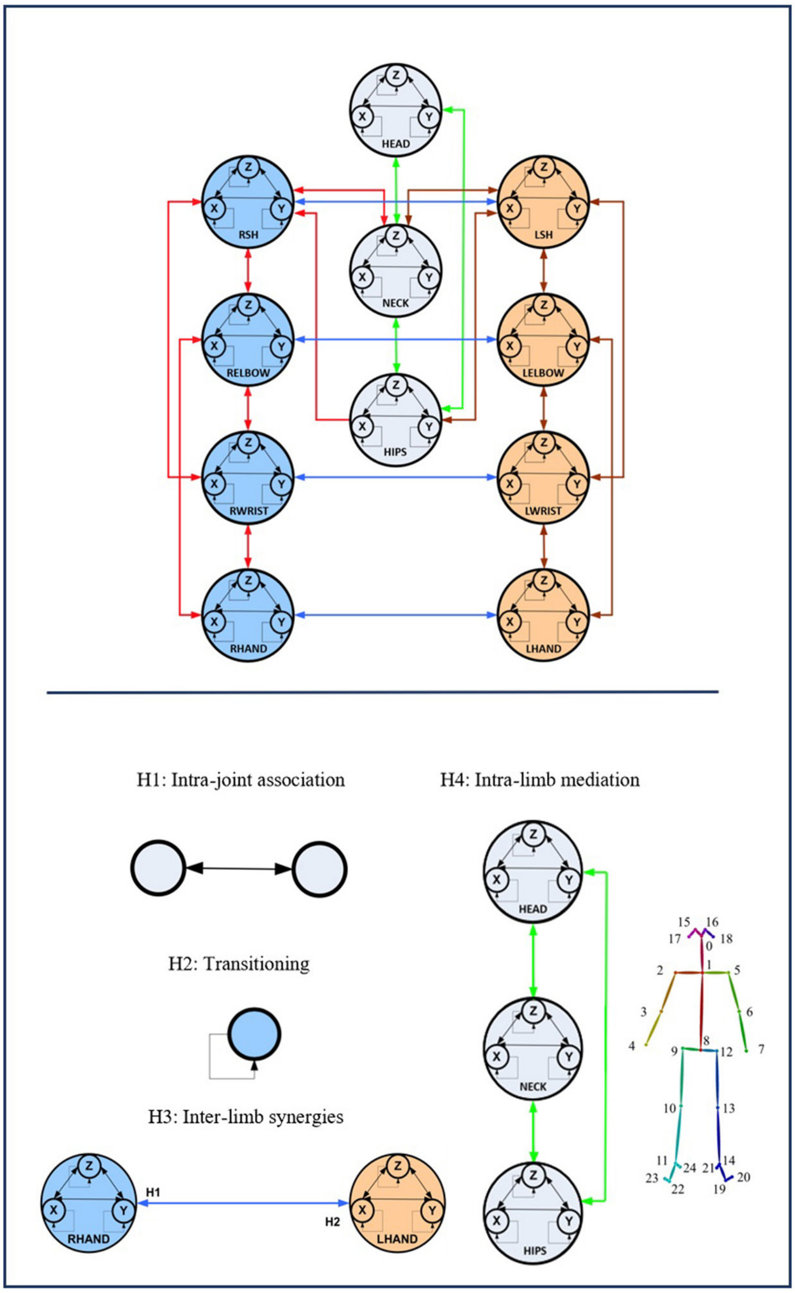


FIGURE 2 | The full-body assumptions of the Gesture Operational Model (GOM) are depicted in the figure. Some relationships happen to be bidirectional, while others not. The relationships of the human body are governed by four different assumptions, intra-joint association, transitioning, inter-limb synergies, and intra-limb mediation. On the down-right of the image, the mapping on body is presented. The numbers in the GOM model, represent the corresponding body part of the joints representation from OpenPose framework. The idea of the GOM was based on a previous article of the authors (Manitsaris et al., 2015b)[1].

in the automotive assembly lines of PSA Peugeot Citroën (PSA Group). A dual-arm robot and the worker are facing each other in order to cooperate for assembling motor hoses. More precisely, for the assembling of the motor hoses, the robot gives to the worker one part from the right and one part from the left claw, and the worker takes two hose parts from the robot, joins them, screws them, and finally places the mounted motor hose in a box. In order for the robot to achieve the appropriate level of perception and move accordingly, it needs to make two specific actions "to take a piece in the right claw" and "to take a piece in the left claw." Then, the worker can screw after the first gesture "to assemble" or can choose to screw later during the last assembly subtask. At the end of the assembly task, the worker puts the assembled piece in a box, which means that a cycle has just ended. Therefore, it is important for the robot to recognize the actions "to assemble" and "to screw" of the worker, so as to give at the correct moment the next motor piece with its arm. Twelve operators have been recorded in GV_4.

The four datasets and vocabularies contain professional gestures performed in different industries and contexts. Important differences may be observed between them though. For example, GV_1, GV_3, and GV_4 involve the manipulation of tools from the operator. Therefore, the distribution of variances alternates between high, for example, when moving for grabbing the tool, and low values, for example, when tools or objects are put on a specific position. In GV_1, despite the fact that the gestures are performed in a predefined space, the operator has a certain degree of variation between different repetitions of the same gesture. Human factors such as the level of experience, fatigue, or even stress influence the way these gestures are performed without necessarily having a direct impact on the final result, which is to place the card on the TV. However, this is not the case of the GV_3, where a high level of technicity and dexterity is required. In GV_3, only low spatial and temporal variations can be accepted. The glass blower performs the gestures with a high repeatability from one repetition to another and successfully reproduces the object with exactly the same specifications, for example, size and diameter. The gestural commands of GV_2 are simpler and ampler. A bigger freedom and variation are thus expected in the way they are performed. In GV_4, the operator is performing actions with a high repeatability. Because the dual-arm robot Motoman SDA20 has been used, the operator, depending on whether he/she is left- or right-handed, has various possibilities for grabbing the parts from the robot and the tools.

In GV_1, although all the gestures are performed by a single user, the different positions of the operator in space in each gesture make it an interesting dataset to work on. Also, this dataset appears to have a lot of noise, and it was an opportunity to examine the reaction of the pose estimation framework to noisy data. The second dataset (GV_2) has multiple users, giving the opportunity to examine how gesture recognition works with a high variation among the performance of the same gestures. In the third gestural vocabulary (GV_3), all gestures have been performed by an expert artist. They are fine movements where hands are cooperating in a synchronous way. Consequently, investigating body parts dependencies in this GV becomes extremely interesting. The fourth gestural vocabulary (GV_4) has a robot involved in the industrial routine.

From an *intraclass* variance point of view, the Root-Mean-Square Error (RMSE) is used to evaluate the datasets. The RMSE allows the measurement of the difference between two times series, and it is defined as shown in Equation (1).

$$\mathrm{RMSE} = \sqrt{\overline{(o_1 - o_2)^2}} \tag{1}$$

where o_1 is one of the iterations of a specific gesture within a gestural vocabulary and o_2 is another iteration of the same gesture, among which the variance is to be examined. A high variance between the iterations of each gesture of GV_2 is noticed, which is the expected result, because this dataset consists of gestures performed by six different users (**Table 1**). The *RMSE* for GV_3 appears to have low *intraclass* variation, as expected, because the gestures are performed by an expert, who is able to repeat them in a very precise way.

TABLE 1 | RMSE between the iterations of the real data of GV_2 and GV_3.

GV_2	$\overline{G_{2,1}}$	$\overline{G_{2,2}}$	$\overline{G_{2,3}}$	$\overline{G_{2,4}}$	$\overline{G_{2,5}}$
RMSE	0.0565	0.0523	0.0556	0.0330	0.0407
GV_3	$\overline{G_{3,1}}$	$\overline{G_{3,2}}$	$\overline{G_{3,3}}$	$\overline{G_{3,4}}$	
RMSE	0.0265	0.0302	0.0489	0.0461	

Pose Estimation and Feature Extraction

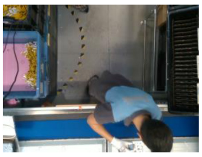
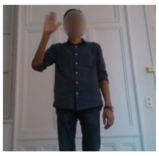
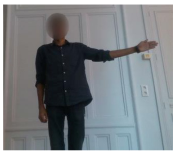
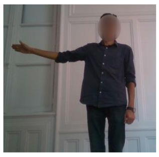
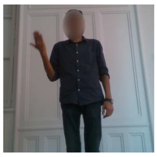
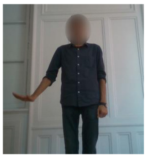
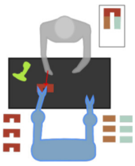
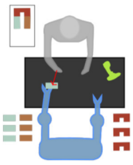
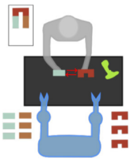
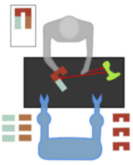
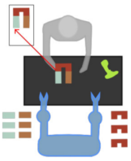
After the motion capturing and recording of the data, each image sequence of the three first datasets is imported to the OpenPose framework, which detects body keypoints on the RGB image and extracts a skeletal model together with the 2D positions of each body joint (Cao et al., 2019) (**Figure 2**). These joints are not necessarily physical joints. They are keypoints on the RGB image, which, in most cases, correspond to physical joint centers. OpenPose uses the neck as the root keypoint to compute all the other body keypoints (or joints). Thus, the motion data are normalized by using the neck as the reference joint. In addition to this, the coordinates of each joint are derived by the width and height of the camera. With regard to GV_4, 3D hand positions are extracted from top-mounted depth imaging by detecting keypoints on the depth map. The keypoints are localized by computing the geodesic distances between the closest body part to the camera (head) and the farthest visible body part (hands), as it is presented in our previous work (Manitsaris et al., 2015a). Any vision-based pose estimation framework may output 2D positions of a low precision, depending on the location of the camera, such as OpenPose for a top-mounted view. However, these errors may not strongly affect the recognition accuracy of our hybrid approach. This is also proved by the fact that our approach outperforms the end-to-end 3DCNNs that does not use any skeletization of the human body to recognize the human actions.

The extracted features for each joint, were the X and Y positions, as they are provided by OpenPose. More specifically, for GV_1, the 2D positions of the two wrists have been used, whereas for GV_2 and GV_3, the 2D positions of the head, neck and shoulder, elbows, wrists, and hands have been used, as they were proven to give optimal recognition results. With regard to GV_4, 3D hand positions are used.

Gesture Operational Model

When a skilled individual performs a professional situated gesture, the whole body is involved combining, thus, theoretical knowledge with practical motor skills. Effective and accompanying body movements are harmonically coordinated to execute a given action. The expertise in the execution of professional gestures is characterized by precision and repeatability, while the body is continuously shifting from one phase to another, for example, from specific postures (small tolerance for spatial variance) to ample movements (high tolerance for spatial variance). For each phase of the movement,

TABLE 2 | Gesture vocabulary of TV assembling dataset, AGV commands dataset, Glassblowing and Human-Robot Collaboration dataset, respectively.

***GV*$_1$ -TV assembly**				
$G_{1,1}$: Take the card from the left side box	$G_{1,2}$: Take the wire from the right-side box	$G_{1,3}$: Connect the wire with the card	$G_{1,4}$: Place the card on the TV chassis	
				
***GV*$_2$-AGV commands**				
$G_{2,1}$: Hello	$G_{2,2}$: Left	$G_{2,3}$: Right	$G_{2,4}$: Speed up	$G_{2,5}$: Speed down
				
***GV*$_3$-Glassblowing**				
$G_{3,1}$: Fix details with pliers	$G_{3,2}$: Tighten base of glass	$G_{3,3}$: Make shape with paper	$G_{3,4}$: Fix shape	
***GV*$_4$-Human-robot collaboration**				
$G_{4,1}$: Take a motor hose part in the robot right claw	$G_{4,2}$: Take a motor hose part in the robot left claw	$G_{4,3}$: Join two parts of the motor hose	$G_{4,4}$: Screw	$G_{4,5}$: Put the final motor hose in a box
				

each body entity, for example, articulation or segment, moves in a multidimensional space over time. When considering the 2D motion descriptors of the movement, two mutually dependent variables represent the entity, for example, X and Y positions. Each of these variables is associated with the other, creating thus a bidirectional relationship between them. Furthermore, they also depend on their history, whereas some entities might "work together" to execute an effective gesture, for example, when an operator assembles two parts. However, a unidirectional dependency might be observed when one entity influences the other entity and not vice versa as well as a bidirectional dependency when both entities influence one each other, for example, when a potter shapes the clay with both hands.

The above observations on situated body movements can be translated into a functional model, which we define here as the GOM, which describes how the body skeletal entities of a skilled individual are organized to deliver a specific result (**Figure 2**). It is assumed that each of the assumptions of "intrajoint association," "transitioning," "intralimb synergies," and "intralimb mediation" contribute at a certain level to the production of the gesture. As far as the *intralimb mediation* is concerned, it can be decomposed into the "interjoint serial mediation" and the "interjoint non-serial mediation." The proposed model works perfectly for all three dimensions (X, Y, and Z), but for reasons of simplicity, it will be presented only for two dimensions, the X and Y. In addition to this, in this work, only positions are used, but the model is designed to be able to receive joint angles as input as well.

Intrajoint Association

It is hypothesized that the motion of each body part (*Entity*) (e.g., right hand) is decomposed in a motion on the X-axis and Y-axis, thus described by two mutually dependent variables. It is assumed that there is a bidirectional relationship between the two variables, defined here as *intrajoint association* and indicated by ●↔●.

Transitioning

It is also assumed that each variable depends on its own history, also called inertia effect. This means that the current value of each variable depends on the values of previous times, also called lag or dynamic effect, which is defined here as *transitioning* and indicated by .

Interlimb Synergies

It is assumed that some body entities work together to achieve certain motion trajectories, for example, hands when assembling two parts, defined here as *interlimb synergies*.

Intralimb Mediation

Interjoint serial mediation

It is assumed that a body entity may depend on its neighboring entities to which it is directly connected to; for example, a glassblower, while using the pipe, moves his/her wrists along with his/her shoulders and elbows. In case this assumption is statistically significant, there is an *interjoint serial mediation*.

Interjoint non-serial mediation

It is assumed that each body entity depends on non-neighboring entities of the same limb; for example, the movement of the wrist may depend on the movement of the elbow and shoulder. Thus, it is highly likely that both direct and indirect dependencies simultaneously occur in the same gesture. Entities are named after the first letters of the respective body joint. More specifically, LSH and RSH represent the left and right shoulders, respectively. Accordingly, LELBOW and RELBOW represent the left and right elbows; LWRIST and RWRIST, the left and right wrists; and LHAND and RHAND, the left and right hands. HEAD, NECK, and HIPS represent, as their names indicate, the head, the neck, and the hips.

So an example of the representation of those assumptions for the X-axis would be as follows:

$$\begin{aligned} Entity_{1,X}\left(t\right) \; = \; & Entity_{1,Y}\left(t-1\right) + Entity_{1,X}(t-1) + \\ & + Entity_{1,X}\left(t-2\right) + Entity_{2,X}\left(t-1\right) \\ & + Entity_{3,X}\left(t-1\right) \end{aligned} \tag{2}$$

Simultaneous Equation System

The simultaneous equation system concatenates the dynamics of an Nth-order system, the GOM, into N first-order differential equations. The number of equations is equal to the number of associated dimensions to a given entity multiplied by the number of body entities. Therefore, the steps to follow are the estimation of the model, with the aim of verifying its structure, as well as the simulation of the model to verify its forecasting ability.

State-Space Representation

The definition of the equations of the system follows the theory of the SS modeling, which gives the possibility for the coefficients to dynamically change over time. An SS model for n-dimensional time series $y(t)$ consists of a *measurement or observation equation* relating the observed data to an m-dimensional state vector $s(t)$ and a Markovian *state or transition equation* that describes the evolution of the state vector over time. The *state equation* depicts the dependence between the system's past and future and must "canalize" through the state vector. The *measurement or observation equation* is the "lens" (signal) through which the hidden state is observed, and it shows the relationship between the system's state, input, and output variables. Representing a dynamic system in an SS form allows the state variables to be incorporated into and estimated along with the observable model.

Therefore, given an input $u(t)$ and a state $s_S(t)$, a SS gives the hidden states that result to an observable output (signal). A general SS representation is as follows:

$$\frac{ds_s}{dt} \; = \; As_S\left(t-1\right) + w\left(t\right) \tag{3}$$

$$y \; = \; C\frac{ds_s}{dt} + Du \tag{4}$$

where Equation (3) is the *state* equation, which is a first-order Markov process; Equation (4) is the *measurement* equation; s_S is the vector of all the state variables; $\frac{ds_s}{dt}$ is the time derivative of the state vector; u is the input vector; y is the output vector; A is the transition matrix that defines the weight of the precedent space; C is the output matrix; and D is the feed-through matrix that describes the direct coupling between u and y; and t indicates time.

When capturing the gestures with motion sensors, Gaussian disturbances (w(t)) are also added in both the state and output equations. After the experiments presented in this work were performed, it was observed that Gaussian disturbances did not change at all the final estimation result, so they were considered to be negligible.

The SS representation of the positions on the X-axis for a body $Entity_{i,j}$*, where i represents the body part modeled in an SS form and j the dimension of each Entity-* according to the GOM structured, as follows:

$$\begin{aligned} \frac{ds_s}{dt} = A{*}s_S\left(t-1\right) \; & = \; \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} Entity_{1,X}\left(t-1\right) \\ -Entity_{1,X}\left(t-2\right) \end{bmatrix} \\ & = \; \begin{bmatrix} \alpha_1 Entity_{1,X}\left(t-1\right) \\ -\alpha_2 Entity_{1,X}\left(t-2\right) \end{bmatrix} \end{aligned} \tag{5}$$

$$\begin{aligned} \overset{(5)}{\Rightarrow} Entity_{1,X}\left(t\right) \; & = \; [1\;1]\frac{ds_s}{dt} + \alpha_3 Entity_{1,Y}(t-1) + \\ & \quad + \alpha_4 Entity_{2,X}\left(t-1\right) \\ & = \; \alpha_1 Entity_{1,X}\left(t-1\right) - \alpha_2 Entity_{1,X}(t-2) + \\ & \quad + \alpha_3 Entity_{1,Y}\left(t-1\right) + \alpha_4 Entity_{2,X}\left(t-1\right) \end{aligned} \tag{6}$$

where α_i are the coefficients that need to be estimated. For simplicity, the inter-joint non-serial mediation is not used in the specific example. In Equation (6), $Entity_X(t-2)$ is subtracted by $Entity_X(t-1)$, indicating the difference between successive levels of dimensions, for example, positions on the Y-axis (transitioning assumption). Equations (5) and (6) occur by Equations (3) and (4), respectively. More specifically, Equation (6) consists of the exogenous variables to which the endogenous ones, coming up from the state equation (Equation 5), are added.

Equation (6) has now the form of a second-order autoregressive (AR) model. An AR model predicts future behavior on the basis of past behavior. The order of the

AR model is adapted in each case according to the data characteristics and the experiments. During the performance of the experiments, the use of an AR model of second order led to better estimation results. As such, in the transitioning assumption, the position values of the two previous time periods (frames) of a given axis are considered.

For the modeling of the full human body, the simultaneous equation system is based on Equations (5) and (6), which consists of two sets of equations for each used entity, one for each dimension X and Y. Thus, for a full-body GOM, we obtain 32 equations describing 32 variables with 64 state variables that contain two endogenous variables for $t-1$ and $t-2$.

As an example, the SS representation for the right wrist is given as follows:

$$\begin{aligned} RWRIST_X\,(t) \;=\; & \alpha_1 RWRIST_X\,(t-1) - \alpha_2 RWRIST_X(t-2) + \\ & + \alpha_3 RWRIST_Y\,(t-1) + \alpha_4 LWRIST_X\,(t-1) \quad (7) \end{aligned}$$

In Equation (7), $RWRIST_X\,(t-1)$ and $RWRIST_X\,(t-2)$ are the endogenous variables, whereas $RWRIST_Y\,(t-1)$, and $LWRIST_X\,(t-1)$ are the exogenous ones.

Computing the Coefficients of the Equations

The coefficients of the simultaneous equation system are computed using the MLE method via Kalman filtering (Holmes, 2016). Let us consider a gesture $G_{j\in\mathbb{N}}$ of a gesture vocabulary $GV_{i\in[1,3]}$ and an observation $\mathcal{O}_{0:k} = \{o_1, \ldots, o_k\}_{k\in\mathbb{N}}$, where o_k is one observation vector and k the total number of observations. Thus, the probability $\mathcal{P}_s$ to observe o_t at time $t \in \left[0,\, k\right]$ will be as follows:

$$\mathcal{P}_s\,(\mathcal{O}_{0:k}) = \prod_{t=0}^{k} \mathcal{P}(o_t\,|\mathcal{O}_{0:t-1}) \qquad (8)$$

where k represents the observed data, $\mathcal{P}\,(o_t|\mathcal{O}_{0:t-1})$ is the probability of o_t given all the observations before time t.

Also, the probability of time series given a set of parameters Ψ is

$$\mathcal{P}\,(\mathcal{O}_{0:t-1}|\Psi) = \prod\nolimits_{t=1}^{k} \exp\left\{-\frac{\left(o_t - \tilde{o}_t^{t-1}\right)^2}{2F_t^{t-1}}\right\}(2\pi|F_t^{t-1}|)^{-\frac{1}{2}}\,d\theta \qquad (9)$$

with variance F_t^{t-1} and mean $\tilde{o}_t^{t-1}$. So the log-likelihood of ψ given data $\mathcal{O}_{0:t-1}$ is

$$\begin{aligned} \log\mathrm{L}\,(\Psi|\mathcal{O}_{0:t-1}) \;=\; & -\frac{k}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{k}\log\left|F_t^{t-1}\right| - \\ & -\frac{1}{2}\sum_{t=1}^{k}\frac{\left(o_t - \tilde{o}_t^{t-1}\right)^2}{F_t^{t-1}} \end{aligned} \qquad (10)$$

For the computation of this log-likelihood, the estimation, the variance, and mean that appear in Equation (4) need to be estimated optimally. Kalman filtering gives the optimal estimates of the mean and covariance for the calculation of the maximum likelihood of ψ. Kalman filtering consists of two main recursive steps, prediction and update. In the first step, there is an estimation of the mean and covariance, along with the predicted error covariance. In the update step, the optimal Kalman gain is computed, so the estimation of mean and covariance from the prediction step is updated according to it. These two steps appear recursively, until the optimal $\tilde{o}_t^{t-1}$ and F_t^{t-1} that fit the observed data are computed. This derives the computation of the coefficients of the SS equations and the forecasting of a new time series given those observed data.

Learning With Hidden Markov Models

HMMs follow the principles of Markov chains that describe stochastic processes. They are commonly used to model and recognize human gestures. They are structured using two different types of probabilities, the transition probability from one state to another and the probability for a state to generate specific observations on the signal (Bakis, 1976). In our case, each professional gesture is associated to an HMM, whereas the intermediate phases of the gesture constitute internal states of the HMM. According to our datasets, these gestures define the gesture vocabulary $GV_{i\in[1,3]} = \left\{G_j\right\}_{j\in\,\mathbb{N}}$.

Let $\mathcal{S}_h$ be a finite space of states, corresponding to all the intermediate phases of a professional gesture. The transition probability is between the states $\mathcal{Q}\,\left(s_h,\, s_h'\right)$, where $s_h,\ s_h' \in \mathcal{S}_h$ are given in the transition matrix $\mathcal{Q} = \left[\mathcal{Q}\left(s_h,\, s_h'\right)\right]$. A hidden sequence of states where $\mathcal{S}_{h0:k} = \{s_{h1}, \ldots, s_{hk}\}_{k\in\mathbb{N}}$, where $s_{hk} \in \mathcal{S}_h$ is also considered. A given sequence of hidden states $S_{h0:k}$ is supposed to generate a sequence of observation vectors $\mathcal{O}_{o:k}$. We assume that the vectors o_k depends only on the state s_{hk}. From now on, the likelihood that the observation o is the result of the state s_h will be defined as $\mathcal{P}_h(o|s_h)$. It is important to outline that in our modeling structure, each internal state of the model depends only on its previous state (first-order Markov property). Consequently, the set of the models for all gestures for every gesture vocabulary is $GV_{j\in[1,3]} = \{HMM_i\}_{i\in\mathbb{N}}$, where $HMM_i = (\varrho_i, \mathcal{Q}_i, \mathcal{P}_{hi})_{i\in\mathbb{N}}$ are the parameters of the model and ϱ_i is the initial state probability. Thus, the recognition becomes an issue of solving three specific problems: *evaluation*, *recognition*, and *learning* (Dymarski, 2011). Each one of those problems was solved with the use of the algorithms, Viterbi (Rabiner, 1989), Baum's "forward" (Baum, 1972), and Baum–Welch, respectively (Dempster et al., 1977).

Gesture Recognition

In the recognition phase, the main goal is to recall with a high precision the hidden sequence of internal states $\mathcal{S}_{h0:k}$ that correspond to the sequences of the observation vectors. Thus, let us consider the observation of motion data $\mathcal{O}_{0:k}$, which need to be recognized. Every HMM_λ with $\lambda \in \left[1, j\right]$ of a given GV_i with $i \in [1,3]$ generates the likelihood $\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_\lambda)$. If there is at least one HMM_ξ with $\xi \in \left[1, j\right]$ that generates $\mathcal{P}_{h\xi} \geq 0.55$, then it is considered that $\mathcal{O}_{0:k}$ is generated by $G_{i,\xi}$. Otherwise, the following quantity is computed for every SS_λ of

GV_i (confidence control):

$$SS_{\lambda}^{score} = \frac{1}{1 + d\left(\mathcal{O}_{0:k}, \mathcal{O}^{s}_{0:k,\lambda}\right)} \tag{11}$$

where d is the minimum distance between the simulated values $\mathcal{O}^{s}_{0:k,\lambda}$ from the model SS_{λ} and the original observations $\mathcal{O}_{0:k}$.

Then, for every SS_{λ} of GV_i, the likelihood $\mathcal{P}'_{h\lambda}\left(\mathcal{O}_{0:k}|HMM_{\lambda}^{SS}\right)$ is computed as follows:

$$\mathcal{P}'_{h\lambda}\left(\mathcal{O}_{0:k}|HMM_{\lambda}^{SS}\right) = \mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_{\lambda}) \cdot SS_{\lambda}^{score} \tag{12}$$

Therefore, the final formula provides the way the algorithm recognizes the observation of motion data $\mathcal{O}_{0:k}$,

$$\mathcal{R}_{GV_i}\left(\mathcal{O}_{0:k}\right) = \begin{cases} \max_1^j\left(\mathcal{P}_{hi}(\mathcal{O}_{0:k}|HMM_i)\right), & \max\left(\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_{\lambda})\right) \geq 0.55 \\ \max_1^j\left(\mathcal{P}'_{h\lambda}\left(\mathcal{O}_{0:k}|HMM_{\lambda}^{SS}\right)\right), & \max\left(\mathcal{P}_{h\lambda}(\mathcal{O}_{0:k}|HMM_{\lambda})\right) < 0.55 \end{cases} \tag{13}$$

EVALUATION

The evaluation of the accuracy and performance of the method follows an "all-shots" approach for the training of the HMMs and a "one-shot" approach for estimating the coefficients of the SS models.

In order to select which gestural iteration to use for computing the coefficients of the SS models, the "leave-one-out method" is used. It is a resampling technique that is also useful for variance and bias estimation (and avoidance), especially when the data are limited. It consists in systematically leaving out one observation from a dataset, calculating the estimator, and then finding the average of these calculations. In our case, the estimator was the likelihood of the HMM when trained with one iteration of a gesture and tested with all the other iterations. The iteration giving the maximum likelihood is selected for computing the coefficients of the SS models.

STATISTICAL SIGNIFICANCE AND SIMULATION OF THE MODELS

In order to evaluate the significance of the assumptions concerning the body part dependencies that are defined within the GOM, a statistical significance analysis is done. The statistical significance p-value indicates whether the assumptions are verified or not. The level of statistical significance is often expressed by using the p-value, which takes values between 0 and 1. Generally, the smaller the p-value, the stronger the evidence that the null hypothesis should be rejected. In this work, the 0.05 p-value was used as the threshold for the statistical significance tests. If the p-value of the estimated coefficient is smaller than 0.05, then the specific coefficient is statistically significant and need to be included in the SS representation of the model.

In the case of the professional gestures, investigating the significance level of the coefficients of each variable within the GOM explains how important is each joint for each gesture in the gestural vocabulary. Examples of some of the gestures from GV_2 and GV_3 are given, to observe cases where some of the coefficients affect strongly the results and need to remain dynamic, whereas others cannot, and can remain constant. In the GOM below, the equation of $G_{2,1}$ for $RWRIST_X$ is as follows, starting from Equation (2).

$$\begin{aligned} RWRIST_X\,(t) &= a_{12}RWRIST_Y\,(t-1) + a_{13}RWRIST_X(t-1) - \\ &\quad - a_{14}RWRIST_X\,(t-2) + a_{15}LWRIST_X\,(t-1) \\ &= \overbrace{-0.0629}^{0.266}\,RWRIST_Y(t-1) + \\ &\quad + \overbrace{1.3438}^{0.00}RWRIST_X(t-1) - \\ &\quad - \left(\overbrace{-0.3648}^{0.00}\right) RWRIST_X(t-2) + \\ &\quad + \left(\overbrace{-0.6625}^{0.449}\right) LWRIST_X(t-1) \end{aligned} \tag{14}$$

Having performed the statistical significance analysis of the model in Equation (14), we get the estimation of the coefficients, where Equation (14) is the general equation for X-axis of the right wrist, along the p-values that indicate the level of significance of each part of the equation. The p-values show that in the case of the $G_{2,1}$, the past values on the same axis appear to be significant, whereas the respective p-values of the left wrist or the Y-axis of the right wrist are not statistically significant. This result was expected, as this gesture is a "hello waving movement," where the right wrist is moving across the X axis and the left wrist remains still through the performance of the gesture, leading to the result that there is no intralimb mediation in this specific gesture.

In the following, there is one more example of the same GV, from gesture $G_{2,3}$, for X-axis (Equation 15) and Y-axis (Equation 16). The numbers above the estimated coefficients correspond to their respective p-values.

$$\begin{aligned} RWRIST_X\,(t) &= a_{12}RWRIST_Y\,(t-1) + a_{13}RWRIST_X(t-1) - \\ &\quad - a_{14}RWRIST_X\,(t-2) + a_{15}LWRIST_X(t-1) \\ &= \overbrace{-0.2871}^{0.00}\,RWRIST_Y\,(t-1) + \overbrace{0.6392}^{0.00} \times \\ &\quad \times RWRIST_X\,(t-1)\,\overbrace{-0.0273}^{0.86}RWRIST_X(t-2) + \\ &\quad + \overbrace{0.0516}^{0.00}\,LWRIST_X(t-1) \end{aligned} \tag{15}$$

$$\begin{aligned} RWRIST_Y\,(t) &= a_{12}RWRIST_X(t-1) + a_{13}RWRIST_Y(t-1) - \\ &\quad - a_{14}RWRIST_Y\,(t-2) + a_{15}LWRIST_Y(t-1) \\ &= \overbrace{-3.9907}^{0.00}\,RWRIST_X(t-1) + \\ &\quad + \overbrace{0.5003}^{0.00}\,RWRIST_Y(t-1) - \end{aligned}$$

$$\begin{aligned}
&-(\overbrace{-0.0818}^{0.616})RWRIST_Y(t-2)+\\
&+(\overbrace{-0.0927}^{0.00})LWRIST_Y(t-1)
\end{aligned} \tag{16}$$

In this gesture, the operator moves his/her right wrist toward his/her right side both on the X and Y axes, indicating to the AGV to turn right. So according to the results, all coefficients appear to be statistically significant, apart from the two previous time-period values of the X-axis of the right wrist. The same results occur for the Y-axis of the same wrist.

To verify the results, a significance level test is presented for $G_{3,2}$ of GV_3. During the performance of this gesture, the glassblower is moving both wrists cooperatively, to tighten the base of the glass piece. The right wrist works more intensively to complete tightening the glass, whereas the left wrist complements the movement by slowly rolling the metal pipe.

$$\begin{aligned}
RWRIST_X\,(t) \;=\;& a_{12}RSH_X\,(t-1)+a_{13}RELBOW_X(t-1)+\\
&+a_{14}RWRIST_Y\,(t-1)+a_{15}LWRIST_X(t-1)+\\
&+a_{16}\,RWRIST_X\,(t-1)-a_{17}RWRIST_X(t-2)\\
=\;&\left(\overbrace{-0.0778}^{0.562}\right)RSH_X\,(t-1)+\overbrace{1.1126}^{0.00}\,RELBOW_X(t-1)+\\
&+\left(\overbrace{-0.4757}^{0.00}\right)\,RWRIST_Y\,(t-1)+\overbrace{0.3423}^{0.00}\times\\
&\times LWRIST_X\,(t-1)+\overbrace{0.4585}^{0.00}\,RWRIST_X(t-1)+\\
&-\overbrace{0.4604}^{0.00}\,RWRIST_X(t-2)
\end{aligned} \tag{17}$$

$$\begin{aligned}
RWRIST_Y\,(t) \;=\;& a_{12}RSH_Y\,(t-1)+a_{13}RELBOW_Y(t-1)+\\
&+a_{14}RWRIST_X(t-1)+a_{15}LWRIST_Y(t-1)+\\
&+a_{16}\,RWRIST_Y\,(t-1)-a_{17}RWRIST_Y\,(t-2)\\
=\;&\overbrace{0.290}^{0.117}\,RSH_Y\,(t-1)+\overbrace{0.3678}^{0.00}RELBOW_Y(t-1)+\\
&+\left(\overbrace{-1.0912}^{0.00}\right)\,RWRIST_X(t-1)+\left(\overbrace{-0.1602}^{0.045}\right)\times\\
&\times LWRIST_Y(t-1)+\overbrace{1.1298}^{0.00}\,RWRIST_Y(t-1)+\\
&-(\overbrace{-0.1679}^{0.00})\,RWRIST_Y(t-2)
\end{aligned} \tag{18}$$

$$\begin{aligned}
LWRIST_X\,(t) \;=\;& a_{12}LSH_X\,(t-1)+a_{13}LELBOW_X(t-1)+\\
&+a_{14}\,LWRIST_Y\,(t-1)+a_{15}RWRIST_X(t-1)+\\
&+a_{16}\,LWRIST_X\,(t-1)-a_{17}LWRIST_X\,(t-2)\\
=\;&\overbrace{0.3668}^{0.00}\,LSH_X\,(t-1)+\overbrace{0.11180}^{0.007}\,LELBOW_X(t-1)+\\
&+\overbrace{0.9589}^{0.00}\,LWRIST_Y\,(t-1)+\left(\overbrace{-0.0126}^{0.339}\right)\times\\
&\times RWRIST_X\,(t-1)+\overbrace{1.1111}^{0.00}LWRIST_X(t-1)+\\
&-(\overbrace{-0.1398}^{0.052})\,LWRIST_X(t-2)
\end{aligned} \tag{19}$$

$$\begin{aligned}
LWRIST_Y\,(t) \;=\;& a_{12}LSH_Y\,(t-1)+a_{13}LELBOW_Y(t-1)+\\
&+a_{14}LWRIST_X\,(t-1)+a_{15}RWRIST_Y(t-1)+\\
&+a_{16}LWRIST_Y\,(t-1)-a_{17}LWRIST_Y\,(t-2)\\
=\;&\overbrace{0.9313}^{0.00}\,LSH_Y\,(t-1)+\overbrace{0.5433}^{0.00}\,LELBOW_Y\,(t-1)\\
&+\overbrace{0.1144}^{0.272}\,LWRIST_X(t-1)+\overbrace{0.0162}^{0.356}\,RWRIST_Y\,(t-1)+\\
&+\overbrace{1.0463}^{0.0}\,LWRIST_Y(t-1)-\overbrace{(-0.1095)}^{0.124}\,LWRIST_Y(t-2)
\end{aligned} \tag{20}$$

In the equations presented above, all coefficients appear to be statistically significant, except from $RSH_X\,(t-1)$ in Equation (17), $LWRIST_X\,(t-1)$ in Equation (18), $LWRIST_X\,(t-1)$ and $RWRIST_X\,(t-2)$ in Equation (19), and $RWRIST_X\,(t-1)$, $RWRIST_Y\,(t-1)$, *and* $LWRIST_X\,(t-1)$ in Equation (20). As a result, the hands of the operator work mostly independently (there appears to be a dependency in the interlimb synergies in Equation 17), whereas all the other assumptions seem to be statistically significant for both X-axis and Y-axis of the right and left wrists.

The simulation of the models is based on the solution of their simultaneous equations system. **Figures 3–5** show examples of the graphical depiction of real observations of motion data together with their simulated values from the SS model of the right wrist. A general conclusion that can be exported by looking at the depictions is that the behavior of the models is very good because the two curves are really close in most cases.

Recognition Accuracy and Comparison with End-to-End Deep Learning Architectures

For the evaluation of the performance and the proposed methodology, the metrics *precision*, *recall*, and *f-score* were calculated. Those metrics are defined as shown below.

$$precision = \frac{\#(true\ positives)}{\#(true\ positives)+\#(false\ positives)} \tag{21}$$

$$recall = \frac{\#(true\ positives)}{\#(true\ positives)+\#(false\ negatives)} \tag{22}$$

Precision, *recall*, and *f-score* are calculated for all the gestures that each gestural vocabulary consists of. For a gesture of class i, $\#(true\ positives)$ represent the number of gestures of class i that were recognized correctly, $\#(false\ positives)$ represent the number of gestures that did not belong in class i, and they were recognized from the algorithm as parts of class i. Finally, $\#(false\ negatives)$ represents the number of gestures belonging to class i that were not recognized as part of it.

More precisely, *precision* represents the rate of gestures that really belong in class i, among those who are recognized as class i, whereas *recall* represents the rate of iterations of gestures of class i that have been recognized as class i. A measure that

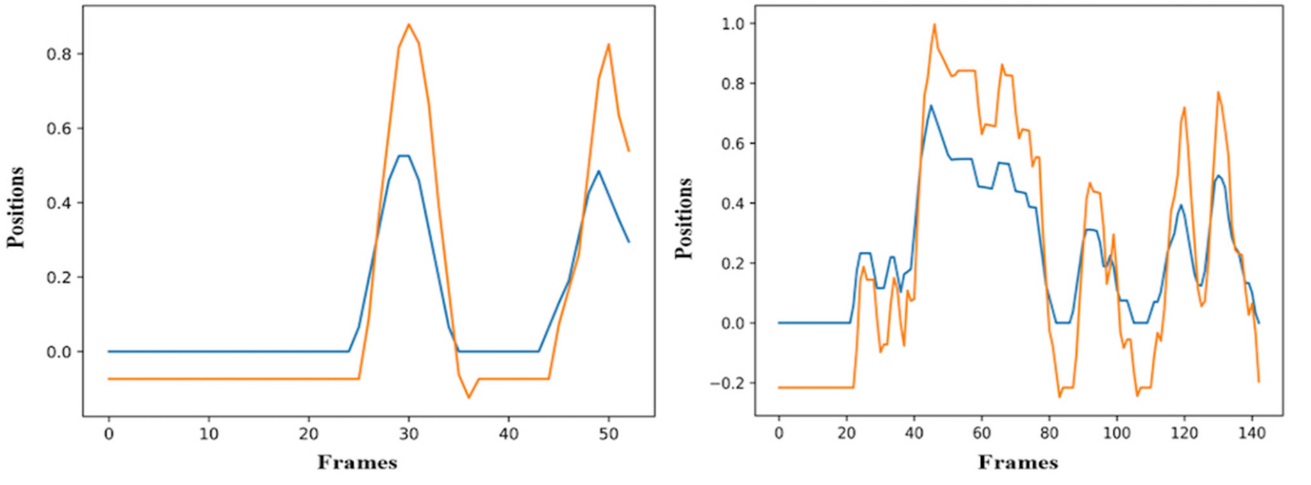


FIGURE 3 | Examples of real motion observations (blue) and simulated values (orange) from the $RHAND_X$ State-Space model of the gesture $G_{1,1}$ (left) and $G_{1,4}$ (right).

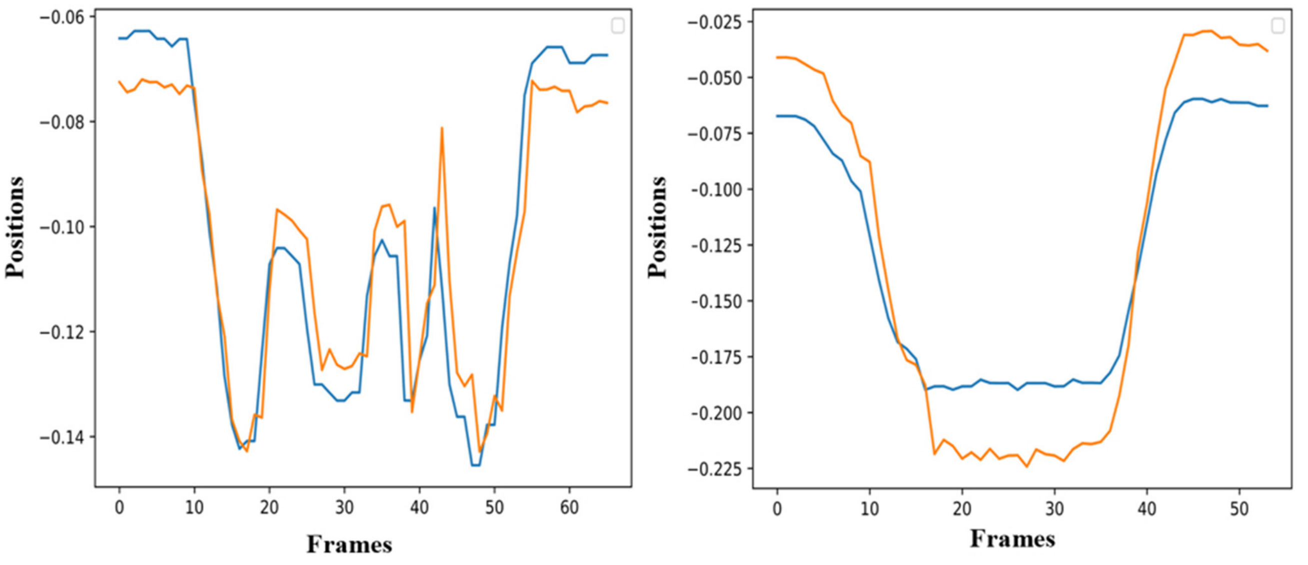


FIGURE 4 | Examples of real motion observations of motion data (blue) and simulated values (orange) from the $RHAND_X$ State-Space model of gesture $G_{2,1}$ (left) and $G_{2,2}$ (right).

combines both precision and recall is the f-score, which is given by Equation (23).

$$f - score = 2\frac{precision{*}recall}{precision + recall} \tag{23}$$

The performance of the algorithms was tested with the four different gestural vocabularies. As presented before, the GV_1 contains four classes, from 44 to 48 repetitions for each. Four hidden states were used for HMM training. To simplify the evaluation task, a simplified GOM with only X and Y positions of two wrists are used for training and recognition. **Table 4** presents the results when only the HMMs are used for recognition without any confidence control and also the results with the confidence control provided by the simulation of the SS models.

It is possible to observe that HMMs provide a recall superior to 90% in three out of four gestures. The $G_{1,3}$ presents the lowest recall of 81.81%, and this can be due to the fact that this is the most complex gesture, where both hands interact more than in the other three gestures. The lowest precision is detected for the $HMM_{1,4}$. When the SS representation and confidence control is used, the *recall* for $G_{1,2}$ is slightly improved, whereas in the case of the $G_{1,3}$, a significant improvement of 15.91% is achieved. Especially for $G_{1,3}$, the improvement can be justified by the fact that the operator is connecting the wire with a very small card outside the conveyor. Thus, the operator has the possibility to perform very small movements in different positions of his/her

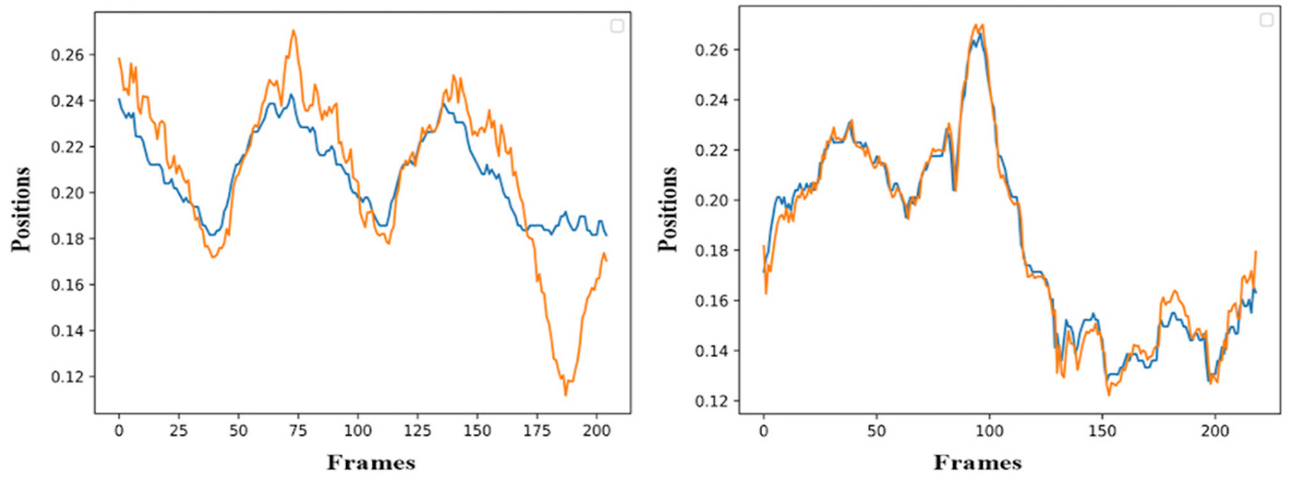


FIGURE 5 | Examples of real motion observations (blue) and simulated values (orange) from the $RHAND_X$ State-Space model of the gesture $G_{3,1}$ (left) and $G_{3,4}$ (right).

TABLE 3 | Confusion matrix using HMM, HMM^{SS}, and $3DCNN$ based on the model of Tran et al. (2015) approaches for GV_1.

	$HMM_{1,1}$	$HMM_{1,2}$	$HMM_{1,3}$	$HMM_{1,4}$	**Recall (%)**
$G_{1,1}$	48	0	0	0	100
$G_{1,2}$	0	44	0	2	95.65
$G_{1,3}$	1	4	36	3	81.81
$G_{1,4}$	0	0	0	46	100
Precision (%)	97.9	91.66	100	90.2	
	$HMM^{SS}_{1,1}$	$HMM^{SS}_{1,2}$	$HMM^{SS}_{1,3}$	$HMM^{SS}_{1,4}$	**Recall (%)**
$G_{1,1}$	47	0	0	1	97.91
$G_{1,2}$	1	45	0	0	97.82
$G_{1,3}$	0	1	43	0	97.72
$G_{1,4}$	0	4	0	42	91.3
Precision (%)	97.91	90	100	97.67	
	$3DCNN_{1,1}$	$3DCNN_{1,2}$	$3DCNN_{1,3}$	$3DCNN_{1,4}$	**Recall (%)**
$G_{1,1}$	48	0	0	0	100
$G_{1,2}$	0	43	0	5	89.5
$G_{1,3}$	0	0	44	0	100
$G_{1,4}$	0	7	0	39	84.7
Precision (%)	100	86	100	88.6	

workplace. The *precision* of $HMM^{SS}_{1,4}$ has been also positively impacted by the SS augmenting the precision from 90.2 to 97.67%. However, a slight decline can also be seen in the case of $G_{1,4}$ recall (**Table 3**).

The GV_2 contains five classes and 16 repetitions of each gesture, and one to 11 hidden states were used for the machine learning gesture recognition engine according to the best states' number for each iteration. The joints selected for training with GV_2 were the wrist, elbow, and shoulder joints for each hand, along with the neck. In **Table 4**, *precision* and *recall* using only HMM and the HMM^{SS} approach is presented, respectively. For $G_{2,1}$, $G_{2,4}$, and $G_{2,5}$, ergodic topology was used, as iterations of the same gestural part appear during the performance of each gesture, whereas left to right topology was used for the rest of the gestures. *Precision* appears improved for every model, whereas *recall* is decreased for $G_{2,2}$ and $G_{2,5}$ (**Table 4**).

GV_3 consists of four different gestures with 35, 34, 21, and 27 repetitions, respectively. 5 to 20 hidden states were used for training the gesture recognition algorithm, the number of which were again computed for every iteration in the resampling phase. The joints selected for training with GV_3 were again the wrist, elbow, and shoulder joints for each hand, along with the neck. *Precision* appears improved in almost every observation and maximum likelihood. The *recall* in almost every gesture has remained stable except from $G_{3,3}$, where it was increased by +4% (**Table 5**).

GV_4 consists of five different gestures. The clusters used in the k-means approach in combination with an HMM with 12 hidden states were 25. The proposed methodology in this work performed better than the rest of the machine learning methods, with f-score results improved by +12% (**Table 6**).

In **Table 7**, the comparison of mean f-scores for each GV and each approach is presented. The score of GV_1 and GV_2 was improved by ∼2%, while the most important contribution is observed for the GV_3. The HMM^{SS} allows to improve significantly (+7.5%) the recognition results of this last dataset.

A similar conclusion can be extracted from the same table, where the total accuracy for the GV_3 has reached 80.34% from 70.94%. The accuracy improvement of the two other datasets remain at the same level with the one of the mean f-score, around +2%.

In order to compare the results of the approach proposed in this paper with other classification techniques, a DL end-to-end 3D CNN has been used to classify the gestures of the three first vocabularies described in *Industrial Datasets and Gesture Vocabularies*. More precisely, a $3DCNN$ has been initially trained on spatiotemporal features from a medium-sized UCF-101 video

TABLE 4 | Confusion matrix using HMM, HMM^{SS}, and $3DCNN$ based on the model of Tran et al. (2015) approach for GV_2.

	$HMM_{2,1}$	$HMM_{2,2}$	$HMM_{2,3}$	$HMM_{2,4}$	$HMM_{2,5}$	**Recall (%)**
$G_{2,1}$	14	1	0	1	0	87.5
$G_{2,2}$	3	13	0	0	0	81.25
$G_{2,3}$	0	0	16	0	0	100
$G_{2,4}$	5	0	0	11	0	68.75
$G_{2,5}$	2	0	0	2	12	75
Precision (%)	58.3	92.8	100	78.5	100	
	$HMM^{SS}_{2,1}$	$HMM^{SS}_{2,2}$	$HMM^{SS}_{2,3}$	$HMM^{SS}_{2,4}$	$HMM^{SS}_{2,5}$	**Recall (%)**
$G_{2,1}$	16	0	0	0	0	100
$G_{2,2}$	4	12	0	0	0	75
$G_{2,3}$	0	0	16	0	0	100
$G_{2,4}$	2	0	0	14	0	87.5
$G_{2,5}$	3	0	0	3	10	62.5
Precision (%)	64	100	100	82.3	100	
	$3DCNN_{2,1}$	$3DCNN_{2,2}$	$3DCNN_{2,3}$	$3DCNN_{2,4}$	$3DCNN_{2,5}$	**Recall (%)**
$G_{2,1}$	16	0	0	0	0	100
$G_{2,2}$	0	16	0	0	0	100
$G_{2,3}$	0	0	16	0	0	100
$G_{2,4}$	0	0	0	12	4	75
$G_{2,5}$	8	0	0	0	8	50
Precision (%)	66.6	100	100	100	66.66	

TABLE 5 | Confusion matrix using HMM, HMM^{SS}, and the Tran et al. (2015) $3DCNN$ approach for GV_3.

	$HMM_{3,1}$	$HMM_{3,2}$	$HMM_{3,3}$	$HMM_{3,4}$	**Recall (%)**
$G_{3,1}$	31	2	1	1	88.57
$G_{3,2}$	0	33	1	0	97.05
$G_{3,3}$	2	2	16	1	76.19
$G_{3,4}$	0	0	0	27	100
Precision (%)	93.93	89.18	88.88	93.1	
	$HMM^{SS}_{3,1}$	$HMM^{SS}_{3,2}$	$HMM^{SS}_{3,3}$	$HMM^{SS}_{3,4}$	**Recall (%)**
$G_{3,1}$	31	2	1	1	88.57
$G_{3,2}$	0	33	1	0	97.05
$G_{3,3}$	1	1	17	2	80.95
$G_{3,4}$	0	0	0	27	100
Precision (%)	96.87	91.66	89.47	90	
	$3DCNN_{3,1}$	$3DCNN_{3,2}$	$3DCNN_{3,3}$	$3DCNN_{3,4}$	**Recall (%)**
$G_{3,1}$	35	0	0	0	100
$G_{3,2}$	0	27	7	0	79.4
$G_{3,3}$	2	2	17	0	80.9
$G_{3,4}$	0	0	0	27	100
Precision (%)	94.5	93.1	70.8	100	

TABLE 6 | Confusion matrix using HMM and HMM^{SS} approach for GV_4.

	$HMM_{4,1}$	$HMM_{4,2}$	$HMM_{4,3}$	$HMM_{4,4}$	$HMM_{4,5}$	**Recall (%)**
$G_{4,1}$	42	1	0	0	1	95.4
$G_{4,2}$	3	86	0	0	1	95.5
$G_{4,3}$	0	0	75	2	12	84.2
$G_{4,4}$	1	0	0	43	0	97.7
$G_{4,5}$	2	0	3	0	75	93.7
Precision (%)	87.5	98.8	96.15	95.5	86.2	
	$HMM^{SS}_{4,1}$	$HMM^{SS}_{4,2}$	$HMM^{SS}_{4,3}$	$HMM^{SS}_{4,4}$	$HMM^{SS}_{4,5}$	**Recall (%)**
$G_{4,1}$	42	1	0	0	1	95.5
$G_{4,2}$	3	86	0	0	1	95.5
$G_{4,3}$	0	0	87	0	2	89.8
$G_{4,4}$	5	2	0	37	0	84
$G_{4,5}$	1	0	5	0	74	92.5
Precision (%)	82.3	96.6	94.5	100	94.8	

dataset, and the pretrained weights have been used to fine-tune the model on small-sized datasets including images of operators performing customized gestures in industrial environments.

The architecture of the network is based on the one proposed in Tran et al. (2015) with four convolution and two pooling layers, one fully connected layer, and a softmax loss layer to predict action labels. It has been trained from scratch on the UCF-101[2] video dataset, using batch size of 32 clips and the Adam optimizer (Kingma and Ba, 2014) for 100 epochs, with the Keras DL framework (Chollet, 2015). The entire network was frozen, and only four last layers were fine-tuned on customized datasets by backpropagation.

[2] https://www.crcv.ucf.edu/data/UCF101.php

TABLE 7 | Comparison of mean *f-scores* and final accuracies of each GV for HMM and HMM^{SS} approach.

Mean f-score	**Datasets**			
	$GV_1\%$	$GV_2\%$	$GV_3\%$	$GV_4\%$
HMM	94.34	83.1	90.64	92.1
HMM^{SS}	96.21	85	91.57	92.29
$k-\text{means}+HMM$	-	-	-	80
$3DCNN$	93.4	84	90	-
Total accuracy	**Datasets**			
	$GV_1\%$	$GV_2\%$	$GV_3\%$	$GV_4\%$
HMM	94.56	82.5	91.45	92.5
HMM^{SS}	96.19	85	92.3	93.94
$k-\text{means}+HMM$				82
$3DCNN$	93.5	87	89	

For the datasets GV_1, GV_2 and GV_3 also the $3DCNN$ results are presented, based on the model of Tran et al. (2015). For GV_4 the results are compared to those presented in Coupeté et al. (2019) using a k-means and HMM approach, with 25 clusters and discrete HMMs with 12 hidden states.

TABLE 8 | Theil inequality coefficient, root mean squared error, for one example of the X coordinate of the right wrist per dataset.

Gestures	**Theil Inequality** ***U***	**Bias proportion** U^B	**Variance proportion** U^V	**Covariance proportion** U^C	**RMSE**
$G_{1,1}$	0.018388	0.009178	0.081456	0.909366	0.028904
$G_{2,1}$	0.0000373	0	0.017247	0.983653	0.007461
$G_{3,1}$	0.0000161	0	0.008713	1.041715	0.003277
$G_{4,1}$	0.010059	0	0.039551	0.960449	0.018053

The comparison of recognition accuracy results between $HMMs$, HMM^{SS}, and $3DCNN$ is shown in **Tables 3**, **7**. As far as the GV_1 is concerned, the use of a $3DCNN$ improves the recognition of only one gesture ($G_{1,1}$) as shown in **Table 3**. However, in total, the HMM^{SS} outperforms the other two methods, reaching a total accuracy of 96.19% (**Table 7**). In the second dataset (GV_2), $3DCNN$ does not achieve a satisfying recognition result for the $G_{2,5}$ (66%) in comparison with other methods that reach 100% (**Table 4**); and in total, HMM^{SS} still performs the best as it is possible to observe in **Table 7**. In the GV_3, the HMM^{SS} performs again the best among the three methods, as shown also in **Table 7**, with a total accuracy almost +4% higher and an f-score of +1.5% higher than the DL method.

Forecasting Ability for Motion Trajectories

For the evaluation of the ability of the four SS models that are used to explain the assumptions of the two-entity GOM, a simulation using Equation (2) for all three dimensions and for all used joints was performed (**Table 8**). It includes the computation of Theil's inequality coefficient (U) and its decomposition into the inequality of bias proportion U^B, variance proportion U^V, and covariance proportion U^C. U^B examines the relationship between the means of the actual values and the forecasts, U^V considers the ability of the forecast to match the variation in the actual series, and U^C captures the residual unsystematic element of the forecast errors (Wheelwright et al., 1997). Thus, $U^B + U^V + U^C = 1$. The Theil inequality coefficient measures how close the simulated variables are to the real variables, and it gets values from 0 to 1. The closer to 0 the value of this factor is, the better the forecasting of the variable. Also, the forecasting ability of the model is better when U^B and U^V are close to 0 and U^C is close to 1. The computed coefficients as shown in **Table 8** and result to a sufficient forecasting performance of the simulated model, and the error results reinforce this conclusion.

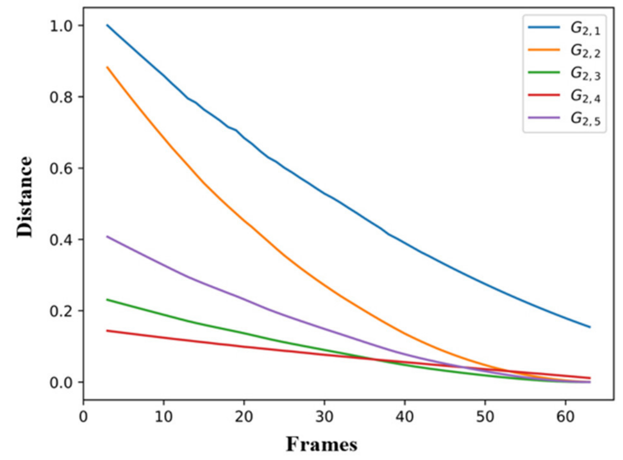


FIGURE 6 | Forecasting performance of all the SS models of GV_2 on the variable $RWRIST_X$ based on $G_{2,4}$.

Also, because the U^V values are really very close to 0, we could extract the potential conclusion that model is able to forecast efficiently even when the real motion data vary significantly (e.g., different operators).

Finally, **Figure 6** presents an example of trajectory forecasting for GV_2. More specifically, the forecasting performance of all the SS models of GV_2 on the variable $RWRIST_X$ is presented, when an unknown observation with data from the $G_{2,1}$ is provided to them. The similarity or distance metric from the DTW is plotted on **Figure 6** taking as input for every time t: 1) the simulated values of the $RWRIST_X$ on $G_{2,4}$ when providing it with real observations until $t = 1$ (starting from $t = 3$), and 2) the real observations between t and the end of the sequence. The distance becomes minimum (high similarity) from the very first time-stamp for the SS model of $G_{2,4}$.

Sensitivity Analysis

As mentioned, the GOM depicts all the possible relationships that take place during the process of the performance of a gesture. Following the GOM, the next steps are the estimation of the model, its dynamic simulation, and its sensitivity analysis. All those steps lead to checking the model's structure, forecasting ability, and its reaction to shocks of its variables, respectively.

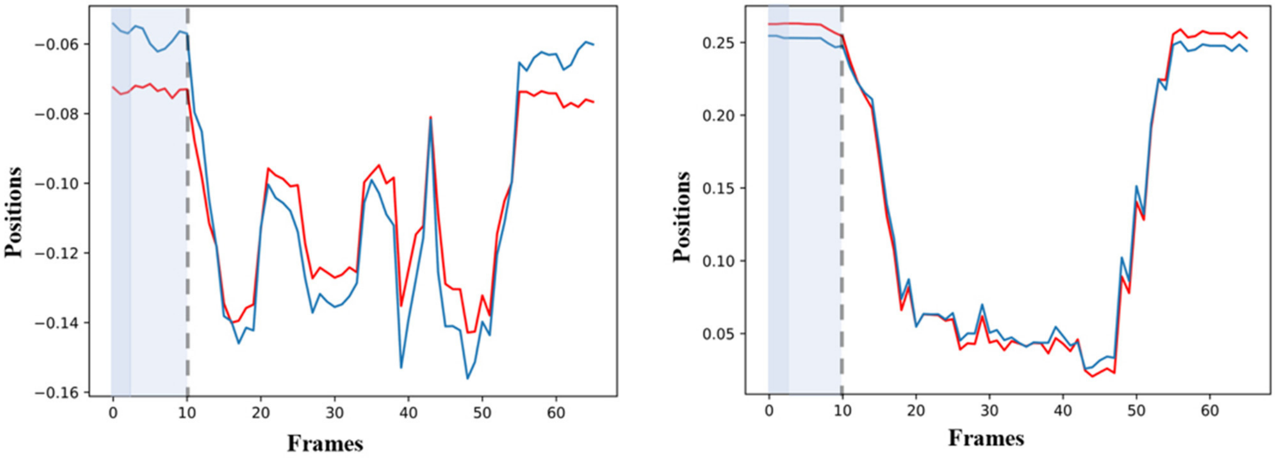


FIGURE 7 | Left: Diagram of the simulated forecasted values of $RWRIST_X$ before the disturbance (red) and simulated forecasted values of $RWRIST_X$ (blue), after the shock on the values of $RWRIST_Y$ by 80% for two frames. Right: Diagram of the simulated forecasted values of $RWRIST_Y$ before the disturbance (red) and simulated forecasted values of $RWRIST_Y$ after the shock on the values of $RWRIST_X$ by 80% for two frames.

The sensitivity analysis of the simulated GOM follows two steps. During the first step, all the simulated values of the model are being estimated, after an artificial shock is provoked for the first two frames. In the second step, all the simulated values that came up after the disturbance are being compared with the simulated values before it (baseline). For example, in **Figure 7**, the simulated values of $RWRIST_X$ are depicted before (red color) and after (blue color) the disturbance on the values of $RWRIST_Y$ by 80%. The disturbance on the simulated variables of $RWRIST_X$ is observed for 10 frames in total, eight more frames than the duration of the initial shock. A similar behavior is also observed for $RWRIST_Y$. The quick adaptation of the models after the application of the artificial shock is observed, which also confirms the low sensitivity of the models to external disturbances.

DISCUSSION

The proposed method for human movement representation on multivariate time series has been used for recognition of professional gestures and forecasting of their trajectories. A comparison has been done between the recognition results of our hybrid approach and the standard continuous HMMs. In general, with both approaches, the best recognition accuracy is achieved for the GV_1. This can be explained by the beneficial *inter*class and *intra*class variations of this vocabulary. The gestures are sufficiently discrete, whereas the different repetitions performed by one operator are sufficiently similar. Nevertheless, we observe an improvement on the recognition accuracy for micro-gestures, when the confidence control of the HMM^{SS} is applied for micro-movements, for example, assembling small pieces, whereas the performance of *HMM*s is satisfactory for macro-movements.

The second-best results are given for the GV_2. Even though these gestures are simpler and do not require any particular dexterity, less good results in recognition accuracy in comparison with the GV_1 are expected mostly because of the high intraclass variation due to multiple users. Although they followed a protocol, each person had significant variations in the way he/she performed the commands. For both datasets, a slight improvement of results has been achieved.

As explained in *Industrial Datasets and Gesture Vocabularies*, the biggest difference of the GV_3 in comparison with the other two gesture vocabularies is the low *inter*class variation because the gestures are similar between them. In three out of four gestures, common gestural patterns are presented: the glass master if controlling the pipe with the left hand is manipulating the glass with the right while sitting, and so forth. These common gestural patterns generate the low *intra*class variation. This low variation can be due to the high level of expert's dexterity, the use of a predefined physical setup (metallic construction) that places his/her body and gestures in a spatial framework (situated gestures) and the use of professional tools that also reduces potential freedom in gesture performances. The low *intra*class variation is also underlined by the comparison of the *RMSE* values for different repetitions of the same gesture performed by the same person. The *HMMs* are thus expected to provide less good results among the four datasets, for the GV_3, because this method may struggle in managing low interclass variation. An important similarity between classes is expected to augment the uncertainty in the maximum likelihood probabilities given by the *HMMs*. This hypothesis can be confirmed through the current recognition results on the basis of *HMMs*. However, it can be clearly noticed that HMM^{SS} had the most beneficial impact on the recognition accuracy of the GV_3. A conclusion can be thus formulated that the proposed methodology permits the improvement of the gesture recognition results to a significant level.

The recognition results of all the three gestural vocabularies using machine learning methods were also compared with those when using *3DCNNs* as a DL method for gesture recognition. In all of the three experiments, the HMM^{SS} method outperformed, and especially in GV_1, achieved +3% higher f-score and accuracy compared with the 3*DCNNs*. In the gestural vocabularies GV_1 and GV_3, the *HMM* method, even if it was not combined with the SS method, achieved slightly higher f-score results than the 3*DCNNs*.

As far as the GV_4 is concerned, our current approach of continuous *HMMs* and SS outperforms our previous one that used *k*-means and discrete *HMMs* (**Table 7**). More precisely, an improvement of at least +12% is observed on the mean f-score, together with an improvement of at least +10% at the total accuracy.

As far as the ability of the models to effectively simulate the professional gestures is concerned, the graphical depiction of the simulated values of the models together with the real motion data can lead to encouraging conclusions. Initially, the simulated values follow very well the real ones for the whole gesture. In particular, the results on GV_1 are quite promising because the pose estimation had some fails because of the top-mounted camera. Nevertheless, the changes or discontinuities on the motion data did not affect the simulation ability of the models. With the regard to the forecasting ability of the models, it is obvious that if the model follows the trajectory from the very beginning, then its forecasting ability is maximized, which is the case in **Figure 6**. Moreover, the evaluation of the forecasting ability of the models using the coefficient of Theil is also encouraging, thus opening a possibility for an efficient forecasting of motion trajectories. In parallel, the sensitivity analysis applied to equations variables proves forecasting's ability of the model to react rapidly to shocks and to provide a solid prediction of motion trajectories.

CONCLUSION AND FUTURE WORK

In this paper, a Gesture Operational Model is proposed to describe how the body parts cooperate to perform a professional gesture. Several assumptions are formulated that determine the dynamic relationship between the body entities within the execution of the human movement. The model is based on the SS statistical representation, and a simultaneous equation system for all the body entities is generated, which is composed of a set of first-order differential equations. The coefficients of the equation system are estimated using the MLE, and its simulation generates a tolerance of the spatial variance of the movement over time. The scientific evidence of the *GOM* is evaluated through its ability to improve the recognition accuracy of gestural time series that are modeled using continuous HMMs. Four datasets have been created for this experiment, corresponding to professional gestures from industrial real-life scenarios. The proposed approach overperformed the recognition accuracy of the *HMMs* by approximately +2% for two datasets, whereas a more significant improvement of +10% has been achieved for the third dataset with strongly situated professional gestures. Furthermore, the approach has been compared with an end-to-end 3DCNN approach, and the mean f-score of the proposed method is significantly higher than the DL, varying approximately from +1.57 to +2.9% better performance, depending on the dataset. A second comparison is done by using a previously recorded industrial dataset from a human–robot collaboration. The proposed approach gives $\sim$ +13% for the mean f-score and +12% for total accuracy, compared with our previous hybrid *k*-means and discrete HMM approach.

Moreover, the system is simulated through the solution of its equations. Its forecasting ability has been evaluated by comparing the similarity between the real and simulated motion data, using at least two real observations to initialize the system, as well as by measuring the Theil inequality coefficient and its decompositions. This paper opened a potential for investigating simultaneous real-time probabilistic gesture and action recognition, as well as forecasting of human motion trajectories for accident prevention and very early detection of the human intention. Therefore, our future work will be focused on extending the proposed methodology for real-time recognition and enhancing the GOM to include kinetic parameters as well. Finally, there will be a continuous enrichment of the datasets by adding new users and more iterations.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available. The datasets generated for this study are anonymous and by the time of the article submission the authors do not have authorization from the industries to publish them. Negotiation is being done with the companies and organizations involved to make the data publicly available.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

SM: Conceptualisation of the methodology and definition of the scientific and industrial needs to address with respect to the human motion analysis, machine learning, and pattern recognition. GS: Contribution to the recording of the dataset, implementation of the experiments, and configuration of the statistical parameters of State-Space. DM: Leading the recording of the datasets, pose estimation and festure selection, developing and integrating the algorithms, and MLE and kalman filtering fine tuning the forecasting ability of the models. AG: Definition of the protocol for the recordings, human factor analysis of the motion data and definition of the vocabularies, and interpretation of the recognition results.

FUNDING

The research leading to these results has received funding by the EU Horizon 2020 Research and Innovation Programme under

Grant Agreement No. 820767, CoLLaboratE project, and Grant No. 822336, Mingei project.

ACKNOWLEDGMENTS

We acknowledge the Arçelik factory as well as Cerfav, the European Centre for Research and Training in Glass Arts—CRT for glass workers and National Innovation Center for Crafts and especially Jean-Pierre Mateus, master glassblower and trainer, for contributing to the motion capturing. Part of this research was supported by the Chair PSA Peugeot Citroën Robotics and Virtual Reality, led by MINES ParisTech and supported by PEUGEOT S.A. The partners of the Chair cannot be held accountable for the content of this paper, which engages the authors' responsibility only.

REFERENCES

Bakis, R. (1976). Continuous speech recognition via centisecond acoustic states. *J. Acoust. Soc. Am.* 59:S97. doi: 10.1121/1.2003011

Baum, L. E. (1972). "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," in *Proceedings of the Third Symposium on Inequalities* (New York, NY).

Bevilacqua, F., Guédy, F., Schnell, N., Fléty, E., and Leroy, N. (2007). "Wireless sensor interface and gesture-follower for music pedagogy," *Proceedings of the NIME'07* (New York, NY), 124–129. doi: 10.1145/1279740.1279762

Bevilacqua, F., Müller, R., and Schnell, N. (2005). "MnM: a Max/MSP mapping toolbox," in *Proceedings of the NIME'05* (Vancouver, BC).

Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., and Rasamimanana, N. (2009). "Continuous real time gesture following and recognition," *Proceedings of the 8th International Conference on Gesture in Embodied Communication and Human-Computer Interaction* (Bielefeld). doi: 10.1007/978-3-642-12553-9_7

Bobick, A. F., and Wilson, A. D. (1997). A state-based approach to the representation and recognition of gesture. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 1325–1337. doi: 10.1109/34.643892

Borl, H. (2018). *Toyota is Bucking the Industrial Automation Trend and Putting Humans Back on the Assembly Line*. Available Online at: https://www.rolandberger.com/en/Point-of-View/Automotive-manufacturing-requires-human-innovation.html (accessed October 08, 2019).

Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2016). "Using convolutional 3D neural networks for user-independent continuous gesture recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)* (Cancun), 49–54. doi: 10.1109/ICPR.2016.7899606

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S. E., and Sheikh, Y. A. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* doi: 10.1109/tpami.2019.2929257

Caramiaux, B. (2015). "Optimising the unexpected: computational design approach in expressive gestural interaction," in *Proceedings of the CHI Workshop on Principles, Techniques and Perspectives on Optimization and HCI* (Seoul).

Caramiaux, B., Montecchio, N., Tanaka, A., and Bevilacqua, F. (2015). Adaptive gesture recognition with variation estimation for interactive systems. *ACM TiiS* 4, 1–34. doi: 10.1145/2643204

Coupeté, E., Moutarde, F., and Manitsaris, S. (2019). "Multi-users online recognition of technical gestures for natural humanŰrobot collaboration in manufacturing" in *Autonomous Robots* 43, 1309–1325.

Chollet, F. (2015). *Keras: Deep Learning Library for Theano and Tensorflow*. Available Online at: https://keras.io (accessed November 25, 2019).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B* 39, 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Devineau, G., Moutarde, F., Xi, W., and Yang, J. (2018). "Deep learning for hand gesture recognition on skeletal data," in *13th IEEE Conference on Automatic Face and Gesture Recognition (FG'2018)* (Xi'An). doi: 10.1109/FG.2018.00025

Dimitropoulos, K., Barmpoutis, P., Kitsikidis, A., and Grammalidis, N. (2016). Classification of multidimensional time-evolving data using histograms of grassmannian points. *IEEE Trans. Circuits Syst. Video Technol.* 28, 892–905. doi: 10.1109/TCSVT.2016.2631719

Duprey, S., Naaim, A., Moissenet, F., Begon, M., and Cheze, L. (2017). Kinematic models of the upper limb joints for multibody kinematics optimisation: an overview. *J. Biomech.* 62, 87–94. doi: 10.1016/j.jbiomech.2016.12.005

Dymarski, P. (2011). *Hidden Markov Models: Theory and Applications* (IntechOpen). doi: 10.5772/601

Fu, Y. (ed.). (2016). *Human Activity Recognition and Prediction.* Switzerland: Springer. doi: 10.1007/978-3-319-27004-3

Holmes, E. E. (2016). *Kalman Filtering for Maximum Likelihood Estimation Given Corrupted Observations*. Seattle: University of Washington.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015* (San Diego, CA: Conference Track Proceedings). Available online at: http://arxiv.org/abs/1412.6980

Lech, M., and Kostek, B. (2012). Hand gesture recognition supported by fuzzy rules and kalman filters. *Int. J. Intell. Inf. Database Syst.* 6, 407–420. doi: 10.1504/IJIIDS.2012.049304

Li, F., Köping, L., Schmitz, S., and Grzegorzek, M. (2017). "Real-time gesture recognition using a particle filtering approach," in *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods. Vol. 1* (Porto: ICPRAM), 394–401. doi: 10.5220/0006189603940401

Li, S. Z., Yu, B., Wu, W., Su, S. Z., and Ji, R. R. (2015). Feature learning based on SAE–PCA network for human gesture recognition in RGBD images. *Neurocomputing* 151, 565–573. doi: 10.1016/j.neucom.2014.06.086

Manitsaris, S., Glushkova, A., Katsouli, E., Manitsaris, A., and Volioti, C. (2015b). "Modelling gestural know-how in pottery based on state-space estimation and system dynamic simulation," in *6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences* (Las Vegas, NV). doi: 10.1016/j.promfg.2015.07.883

Manitsaris, S., Moutarde, F., and Coupeté, E. (2015a). "Gesture recognition using a depth camera for human robot collaboration on assembly line," in *ScienceDirect 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated* Las Vegas, NV: Conferences (AHFE).

Mathe, E., Mitsou, A., Spyrou, E., and Mylonas, P. (2018). "Arm gesture recognition using a convolutional neural network, in *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (Zaragoza), 37–42. doi: 10.1109/SMAP.2018.8501886

Molchanov, P., Gupta, S., Kim, K., and Kautz, J. (2015). "Hand gesture recognition with 3D convolutional neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Boston, MA: IEEE), 1–7. doi: 10.1109/CVPRW.2015.7301342

Oyedotun, O. K., and Khashman, A. (2017). Deep learning in vision-based static hand gesture recognition. *Neural Comput. Appl.* 28, 3941–3951. doi: 10.1007/s00521-016-2294-8

Pedersoli, F., Benini, S., Adami, N., and Leonardi, R. (2014). XKin: an open source framework for hand pose and gesture recognition using kinect. *Vis. Comput.* 30, 1107–1122. doi: 10.1007/s00371-014-0921-x

Psarrou, A., Gong, S., and Walter, M. (2002). Recognition of human gestures and behaviour based on motion trajectories. *Image Vis. Comput.* 20, 349–358.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–285. doi: 10.1109/5.18626

Shahroudy, A., Liu, J., Ng, T. T., and Wang, G. (2016). "Ntu rgb+ d: a large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 1010–1019. doi: 10.1109/CVPR.2016.115

Sideridis, V., Zacharakis, A., Tzagkaraki, G., and Papadopouli, M. (2019). "GestureKeeper: gesture recognition for controlling devices in IoT environments," *27th European Signal Processing Conference (EUSIPCO)* (A Coruña: IEEE). doi: 10.23919/EUSIPCO.2019.8903044

Simonyan, K., and Zisserman, A. (2014). "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 1*. (Montreal: MIT Press), 568–576.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference On Computer Vision* (Las Condes), 4489–4497. doi: 10.1109/ICCV.2015.510

Vaitkevičius, A., Taroza, M., BlaŽauskas, T., Damaševičius, R., Maskeliunas, R., and Wozniak, M. (2019). Recognition of American sign language gestures in a virtual reality using leap motion. *Appl. Sci.* 9:445. doi: 10.3390/app9030445

Volioti, C., Manitsaris, S., Katsouli, E., and Manitsaris, A. (2016). "x2Gesture: how machines could learn expressive gesture variations of expert musicians," in *Proceeding of the 16th International Conference on New Interfaces for Musical Expression.*

Wheelwright, S., Makridakis, S., and Hyndman, R. J. (1997). *Forecasting: Methods and Applications.* 3rd edition. New York, NY: Wiley.

Williamson, J., and Murray-Smith, R. (2002). *Audio Feedback for Gesture Recognition* Technical report TR-2002-127, Dept. Computing Science, University of Glasgow.

Yan, S., Xiong, Y., and Lin, D. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence* New Orleans, LA.

Yang R., and Sarkar, S. (2006). "Gesture recognition using hidden markov models from fragmented observations," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (New York, NY), 766–773. doi: 10.1109/CVPR.2006.126

Yang, H. D., Park, A. Y., and Lee, S. W. (2007). Gesture spotting and recognition for human–robot interaction. *IEEE Trans. Robot*. 23, 256–270. doi: 10.1109/TRO.2006.889491

Zalmai, N., Kaeslin, C., Bruderer, L., Neff, S., and Loeliger, H. A. (2015). "Gesture recognition from magnetic field measurements using a bank of linear state space models and local likelihood filtering," in *IEEE 40th International Conference on Acoustics, Speech and Signal Processing* (Brisbane, QLD: ICASSP), 19–24. doi: 10.1109/ICASSP.2015.7178435

Zatsiorsky, V. (2000). *Biomechanics in Sport: Performance Enhancement and Injury Prevention.* Blackwell Publishing; International Olympic Committee.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Noname manuscript No.
(will be inserted by the editor)

Multi-users online recognition of technical gestures for natural Human-Robot Collaboration in manufacturing

Eva Coupeté · Fabien Moutarde · Sotiris Manitsaris

Received: date / Accepted: date

Abstract Human-Robot Collaboration in industrial context requires a smooth, natural and efficient coordination between robot and human operators. The approach we propose to achieve this goal is to use online recognition of technical gestures. In this paper, we present together, and analyze, parameterize and evaluate much more thoroughly, three findings previously unveiled separately by us in several conference presentations: 1/ we show on a *real prototype* that multi-users continuous real-time recognition of technical gestures on an assembly-line is feasible ($\approx 90\%$ recall and precision in our case-study), *using only non-intrusive sensors* (depth-camera with a top-view, plus inertial sensors *placed on tools*); 2/ we formulate an end-to-end methodology for designing and developing such a system; 3/ we propose a method for *adapting to new users* our gesture recognition. Furthermore we present here two new findings: 1/ by comparing recognition performances using several sets of features, we highlight the importance of choosing features that *focus on the effective part of gestures*, i.e. usually *hands movements*; 2/ we obtain new results suggesting that enriching a multi-users training set can lead to higher precision than using a separate training dataset for each operator.

Keywords Real-time online gesture recognition, Human-Robot Collaboration, Multi-users technical gesture recognition, Collaborative robotics, Gesture recognition from depth-video

Center for Robotics,
MINES ParisTech,
PSL Research University
60 boulevard Saint Michel
75006 Paris
France
E-mail: firstname.lastname@mines-paristech.fr

1 Introduction

The development of robots is currently increasing in our society, and also in our industries. Social robots have already been used in various contexts: assistant for elderly people, stimulation for autistic children, guide in museum or sale assistants in stores.

In the industrial context, robots are present since the 1950's. Until recently, they were always located in isolated areas where operators are not allowed to go while robots are running. In the last years, collaborative robots emerged on assembly-lines. These robots are smaller and can work in co-presence or in collaboration with operators, in the same area. Issues arise with the introduction of these robots. The first one is to guarantee the safety of the operators working nearby those collaborative robots. Technology advances allowed to develop "safe" robots, i.e. robots with with a limited strength and embedded sensors to prevent any injury to operators. A second issue is to make the collaboration smooth and efficient between these robots and operators. In this study, we propose to use online recognition of technical gestures to address the issue. We think that gesture recognition can help the robot to synchronize its tasks with the actions of operators, can allow the robot to adapt its speed, and also can make it able to understand if something unexpected happens. In this paper, we use the word "gesture" to refer to the actions needed to perform the tasks on the assembly-line.

The rest of this article is composed of five sections. In Section 2, we present related works, as well as our own previous research, on Human-Robot Collaboration in manufacturing and gesture recognition. In Section 3 we describe our real prototype and use-case, and how to choose the gesture classes that should be recognized for ensuring human-robot coordination. In Section 4, we summarize our end-to-end methodology to *continuously* recognize technical gestures *in real-time*, including a method to adapt to a new user our learnt gesture recognition system. All our experimental results are presented in Section 5. This section also includes in 5.2 new results highlighting the importance of choosing features that *focus on the effective part of gestures* i.e. usually *hands movements*, and in 5.5.2 new comparative evaluations suggesting that enriching a multi-users training set can lead to higher precision than using a separate training dataset for each operator. Finally, we recapitulate and conclude in Section 6.

2 Related and previous works

2.1 Human-Robot Collaboration in manufacturing

The first robots useful for men have been introduced in factories in the 1950's. These robots were able to perform repetitive, tiresome, and dangerous tasks. Since then, industrial robots have been very present on assembly-lines, working on specific areas, away from human operators. Although these robots are efficient, they make assembly-lines not very flexible, and cannot be used on assembling tasks where human presence is required. Nowadays, manufacturers tend to create mixed environments, where robots and operators can work on the same area. This new way of working combines human skills (intelligence, adaptability and dexterity) with robot skills (strength and repeatability). The introduction of collaborative robots in factories provides more flexibility and productivity (Hägele et al, 2002), and relieves human operators from physically-demanding tasks and/or from working using undesirable postures, that can lead to musculo-skeletal disorders. These robots are designed to be intrinsically safe: their strength is limited, and they have built-in sensors which prevent them to hurt operators which are nearby.

Sharing work between an operator and a robot can be executed in different ways. They can work in collaboration on the same task, or on two different tasks in the same area, in co-presence. Shi et al (2012) proposed different degrees of work sharing. At the lowest level, robot and operator do not have any contact and work in two different spaces, but without any barriers between them. On the second level, the operator can go into the robot space, but this will automatically halt it. Finally, in the upper level, the robot and the operator cooperate on a common task. Other studies have been done to prove the feasibility of collaborative tasks with a robot, e.g. the assembly and disassembly of pieces (Corrales Ramón et al, 2012), and the assembly of constant-velocity joint (Cherubini et al, 2016).

However, sharing work between an operator and a robot requires an adaptation from the operator. Human-robot collaboration is not as natural as between humans, and new ways of communication must be established. Hoffman and Breazeal (2007), have shown that the anticipation on a future task can improve the efficiency and fluidity of a human-robot collaboration. Dragan et al (2015), have demonstrated that legible motions from the robot during the execution of a known task enable a more fluent collaboration with a human. (Chen et al, 2015) proposed an approach for recognizing hand gestures of a human operator during an assembly task in collaboration with a robot co-worker. Schrempf et al (2005) proposed a method to synchronize robot and human actions using a Dynamic Bayesian Network. Rickert et al (2007) presented a collaborative robot that is equipped with speech recognition and visual object recognition, and is able to follow the operator hands. This robot uses these informations to anticipate on the next task. Bannat et al (2011) introduced the term “Cognitive Factory” for industrial environments with cognitive capacities, in order to make the machines more autonomous. Lenz et al (2008) created a smart collaborative workspace with several sensors to enable the collaborative robot to understand their environment.

2.2 Gesture recognition

Gesture recognition consists in capturing and interpreting human movements, allowing to understand which action is being performed. It is a growing research field, in which new technologies recently brought significant progresses. Indeed, new sensors (like depth-cameras or light and small inertial sensors) now enable an easy and more complete capture of gestures. In the following parts, we review and describe the usual successive steps needed for creating a gesture recognition system.

2.2.1 Motion capture sensors

Different types of sensors have been used to recognize gestures. The oldest, and historically most used, are RGB cameras. With these cameras, it is possible to have an almost complete (if few occlusions) understanding of the scene and to be non-intrusive. Laptev and Lindeberg (2003), Dollar et al (2005) and Oikonomopoulos et al (2005) proposed methods to detect interest points in RGB videos, and use them to describe the filmed action. Wang et al (2009) proposed to use trajectories of sampling points in successive frames to describe an action.

Depth-cameras are more recent, but already commonly used to recognize human actions, because of 3D information they convey, which makes extraction of gestures easier. Chen et al (2013) proposed a survey on motion analysis using depth-data, Zhang and Parker (2011) adapted to depth-video the cuboid RGB video features, Biswas and Basu (2011) used movement of a person filmed with a fixed depth-camera to recognize gestures.

Inertial sensors are also used for gesture recognition, but are intrusive because they must be fixed onto the user for capturing his/her movements. Bulling et al (2014) proposed an "*Activity Recognition Chain*" to recognize gestures with inertial sensors. Dong et al (2007) and Junker et al (2008) used accelerometers to recognize actions.

Data from several types of sensors can be used simultaneously to recognize gestures. Chen et al (2016) used a depth-camera and a wearable inertial sensor to recognize actions. To fuse the data coming from the different sensors, a decision level scheme was adopted.

2.2.2 Computing features from sensors

Depending on the sensor that is used for capture of movements, different types of features can be extracted. In this section we focus on the extraction of features *from depth-images.*

A first group of features are those related to the global posture of the human, for example his skeleton. Shotton et al (2011) used a large database, composed of real and synthetic images maps, to learn a random decision forest which is then able, using depth differences between pairs of pixels in the depth-map, to establish for each body pixel to which body part it belongs. Schwarz et al (2012) proposed another method to find the skeleton of a person filmed with a depth-camera, but which does not require pre-training on a large database. They compute geodesic distances of each point of the body part to the gravity center. Knowing the standard structure of a human body, they estimate locations of the body joints. Another group of methods consists in finding body parts in depth-map without using any global information on the user's posture. Migniot and Ababsa (2013) use particles filtering with a top-view depth-camera to determine the position of a top human body. However, detection of hands in depth-videos is still challenging because of the usually rather low resolution of these sensors: only hands which are close enough to the camera can be easily segmented. Chen et al (2011) track the hands' location and segment them using a region-growing algorithm. Hamester et al (2013) detect hands in depth images based on Fourier descriptors of contours classified using a SVM. Joo et al (2014) use boosting of depth-difference features for detecting hands in depth images.

2.2.3 Machine-learning algorithm for classification of gestures

Several machine-learning techniques have been used in order to train a system to recognize human gestures. SVMs (Support Vector Machines), HMMs (Hidden Markov Models), and DTW (Dynamic Time Warping) have been widely used.

SVMs enable to optimize the separation boundaries between different classes in a feature space. Ke et al (2007), Bregonzio et al (2009) and Schuldt et al (2004) used SVMs to recognize actions in video.

Instead of using a fixed-size temporal window represented as a vector, HMMs can process inputs as a flow of successive values. Furthermore, they are able to recognize gestures independently from their temporal duration. Yamato et al (1992) used HMMs to recognize gestures in video. Xia et al (2012) recognized gestures using the skeletons extracted from depth-video with the method of (Shotton et al, 2011). Zhu and Pun (2012) also used depth-images and HMMs to recognize gestures. They track the locations of hands, and use their trajectories to recognize the gestures performed. Calinon and Billard (2004) used HMMs to learn gesture from demonstration. Aarno and Kragic (2008) proposed a Layered Hidden Markov Model (LHMM) to model human skills and classify motions into basic action primitives.

DTW is actually a method for time-series alignment and similarity measure. For gesture recognition, it is generally used first for selecting for each class a single most representative template gesture. DTW similarities with these templates can then be combined with any similarity-based classification algorithm (a simple Nearest Neighbor method in many case) for predicting class of an unlabbeled gesture. DTW has been used for instance by (Liu et al, 2009) to recognize actions based on output of accelerometers worn by users. Sempena et al (2011) similarly recognize actions with DTW, but applied to 3D joint angles time evolutions estimated by Kinect built-in skeletization. Reyes et al (2011) have shown that recognition performance by this method can be greatly improved by weighting differently each joint angle depending on its impact on executed gesture.

Other methods are also used. Luo et al (2013) classified actions with a Bag-of-Visual-Words framework. More recently, deep Convolutional Neural Networks approach was adapted to recognize actions in depth-maps: Wang et al (2016) used weighted hierarchical depth motion maps and three-channel deep convolutional neural networks to recognize actions with a small training dataset.

2.3 Our previous work

We have been working since 2012 on technical gesture recognition for collaborative robotics in factories. All our research is conducted on *real prototype "cells" of factory collaborative robotics* developed by french automaker PSA (see Acknowledgements). After a feasibility study using inertial sensors worn by operators (Coupeté et al, 2014), we have conducted a first experimentation of a less intrusive approach: using only a top-viewing depth-camera for capture of gestures (Coupeté et al, 2015). We have then highlighted in (Coupeté et al, 2016b) the significant recognition rate improvement achievable by complementing gesture capture from depth-camera with data from inertial sensors *placed on tools*. Finally in (Coupeté et al, 2016a) we began investigating the multi-users issue, and proposed a simple but efficient way of adapting our gesture recognition module to new operators.

In this paper, we put together all our methods and algorithms recalled above as a proposed generic methodology and pipeline for technical gesture recognition. Furthermore, we compare several feature sets (hands positions only vs. idem + arms postures, etc...), which we had not done in our previous work, and show that best results are obtained with descriptors related only to the *effective* part of gestures (i.e. hands movements). We also conduct new comparative study on user adaptation providing new results suggesting that enriching a multi-users training set can lead to higher precision than using a separate training dataset for each operator.

3 Our Human-Robot Collaboration prototype

We work on a real-life scenario where the worker and the robot share the same workspace and cooperate. The task is inspired from the assembly of motor hoses on production-line supplies preparation. Presently, the assembly process of motor hoses has some drawbacks: the worker has to find the appropriate parts of motor hoses among other motor parts, which is a lack of time and increase the cognitive load of the worker. In our set-up, a dual-arm robot and the worker are facing each other, with a table separating them, see Figure 1. More details on this *real* prototype are given below, and in (Coupeté et al, 2015) and (Coupeté et al, 2016b).

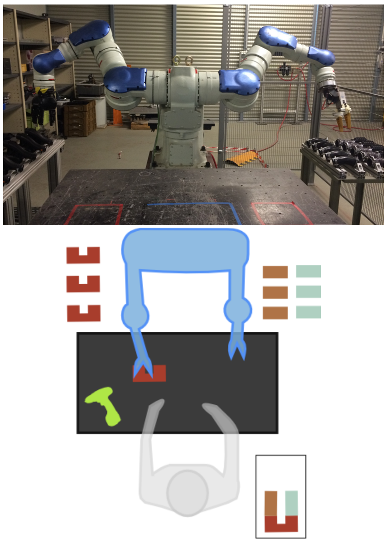


Fig. 1 On top, our human-robot collaboration prototype. On bottom, schematic description of our use-case: an operator is standing in front of a table, taking and assembling parts that are "handed" to him/her by a robot placed on the opposite side of the table.

On an assembly-line, the mounting operations must be executed quickly through a rather strictly-defined succession of elementary and standardized sub-tasks. To ensure a *natural* human-robot collaboration, the robot has to perform its actions according to the task that the operator is currently executing, in order to be useful at the right time, without delaying the worker. In our use-case, the assembling of motor hoses requires the worker to take two hose parts respectively from left and right claw of the robot, join them, screw them, take a third part from the right claw, join it, screw it, and finally place the mounted motor hose in a box. The only actions performed by the robot are giving a piece with the right claw and giving a piece with the left claw. The order of these sub-tasks and how the robot and the operator should be coordinated is presented Figure 2. Such an analysis of the human-robot collaborative work is essential to determine the gesture types that the robot needs to recognize.

In order for the robot to be properly synchronized with the human, it should be able to recognize several gesture classes, that can be deduced from Figure 2. The first two are "to take a piece in the right claw" and "to take a piece in the left claw". The operator can screw after the first gesture "to assemble", or can chose to screw later during the last assembly sub-task. Therefore, the robot should be able to recognize "to assemble" and "to screw", so as to give at the correct moment the third motor piece with its right arm. Finally, at the end of the assembly task, the operator puts the assembled piece in a box, so it is interesting to rec-

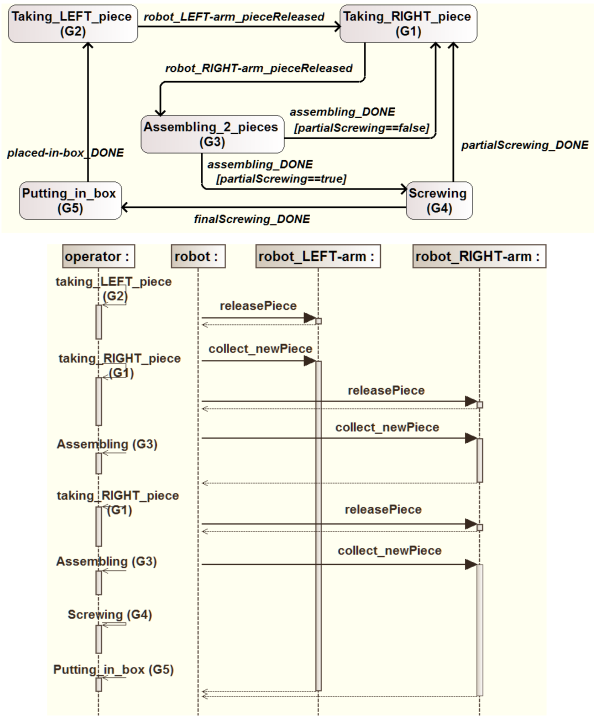


Fig. 2 Analysis of required coordination between Human and Robot: on top, state-transition diagram for the operator tasks; on bottom, sequence diagram of operator-robot interactions.

ognize this gesture in order to understand that a cycle has just ended.

The set of gestures classes to be recognized by our system is therefore rather straightforwardly deduced from above-mentioned sub-tasks as:

1. to take a motor hose part in the robot right claw (G1)
2. to take a motor hose part in the robot left claw (G2)
3. to join two parts of the motor hose (G3)
4. to screw (G4)
5. to put the final motor hose in a box (G5)

Note that in this set-up, the operator chooses the pace during the execution of his sub-tasks, and the robot adapts to it.

4 Methodology for recognition of technical gestures

In this section, we detail our end-to-end methodology for *online* recognition of technical gestures *in real-time*. In the first part 4.1, we present our pipeline (improved and more general than our first versions already presented in (Coupeté et al, 2015) and (Coupeté et al, 2016b)) to recognize gestures, from extraction of features to the classification algorithm. In part 4.2 we describe the two criteria we use to evaluate our gesture recognition system. In part 4.3, we explain how we equipped the scene with an inertial sensor on a tool, and how we refine output from gesture recognition by taking into account the tool-movement information. Finally, in part 4.4, we propose an approach for *adapting to a new user* our system, by limited enrichment of the training database.

4.1 Gesture recognition pipeline

In order to capture the gestures of the operator, we decided to use non-intrusive sensors, for avoiding any discomfort to the operators. Moreover, we want to monitor relative positions of the operator and the robot, while capturing all the operator's movement without potential occlusions. For these reasons, we chose to use a *depth-camera*, filming with a *top-view* for capturing the scene without occlusion. Note that visits on real assembly-lines, and preliminary work on another prototype use-case (door-elements mounting on a continuous line) have convinced us that this choice of sensor type and viewpoint configuration is transferable to most future human-robot collaborative assembly areas. In this section, we explain how we extract body-movement information from the depth-camera.

4.1.1 Posture estimation from depth images

A depth-camera provides information about *3D geometry* of the scene: the value of each pixel corresponds to the distance between the camera and the filmed object to which the pixel belongs.

In many related work on gesture recognition using depth-cameras, input features are simply the successive states of the global human skeleton posture estimation provided by APIs of Kinect for horizontal viewpoint. In our approach using a vertical viewpoint from the top, it was not possible. We therefore needed to extract upper-body (in particular hands) movements from the raw depth-video. We make the assumption that, from top viewpoint, the farthest upper-body parts from the top of the head, using a geodesic measurement, are the two hands. Based on this hypothesis, we have designed an algorithm to locate the operators' hands and estimate arms' postures in the depth-map.

Our algorithm, inspired but significantly modified from (Schwarz et al, 2012) (in which Schwarz et al. estimate global posture, but *only for facing horizontal viewpoint*), is based on estimation of geodesics on body 3D surface. First, we create a 2D graph of the upper-body of the person filmed. Each pixel of this graph are connected with its eight neighbors. We associate for each connection a weight equal to the depth difference between the two pixels, i.e. the difference of the two pixels values. Then, we use Dijkstra algorithm (Dijkstra, 1959) to compute the geodesic distance between each pixel of the upper-body and the top of the head. We are thereby able to detect the two hands positions, and also obtain the geodesically shortest paths between each hand and the top of the head's (which can be used as approximations of arms' postures). Figure 3 illustrates this algorithm, and we refer readers to (Coupeté et al, 2015) for further details on our algorithm. One advantage of this approach for localization of hands is that it is relatively immune to hand occlusion: even when one hand is occasionally hidden from the depth camera (e.g. by another arm), the forearm and arm for this hand are generally still visible, so the geodesic from the head is still properly found and just stops around wrist instead of hand; therefore with this method, hand occlusion leads to only slightly erroneous hand position rather than to absence of information in the data stream.

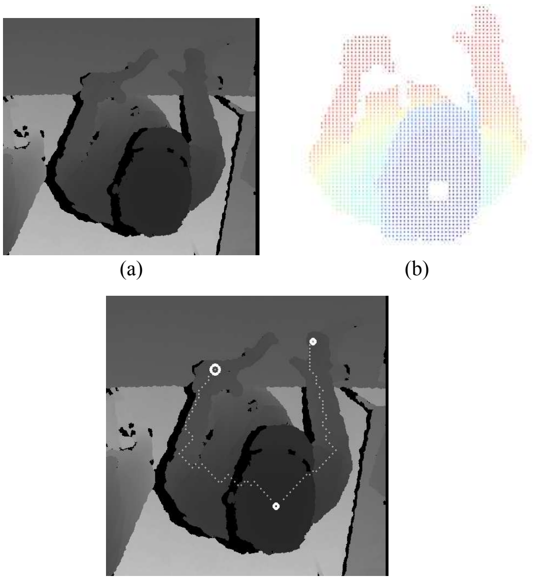

(a)
(b)
(c)

Fig. 3 Our hands localization and upper-body posture estimation algorithm. (a): depth-map from the camera filming an operator with a top-view; (b): geodesic distance for each pixel of the upper-body to the head's top; (c): estimation of the head and two hands locations, plus the geodesically-shortest paths between hands and the head's top (all produced by our algorithm).

4.1.2 Choice of features

In most machine-learning and pattern-recognition tasks, the attainable classification acuracy is strongly depending on the choice of features extracted from raw data and fed into the algorithm. In gesture recognition, it is rather natural and often adopted to use the estimated movements of body parts and joints as features. However, not all of them are equally important, depending on what gesture types should be recognized. Furthermore, it is well-known that inclusion of irrelevant features can reduce recognition rate, either by just adding "noise" to the machine-learning input, or worse by introducing spurious correlations. It is therefore highly recommended to either perform preliminary features selection, or at least to compare recognition performances attained with various sets of human-body related features.

In our case, as exposed above, our dedicated depth-image processing algorithm provides as output:

- 3D location of the head's top;
- estimated 3D locations of the two hands;
- the two 3D geodesically-shortest paths from head to each hand (providing rough approximations of approximations of arms' postures).

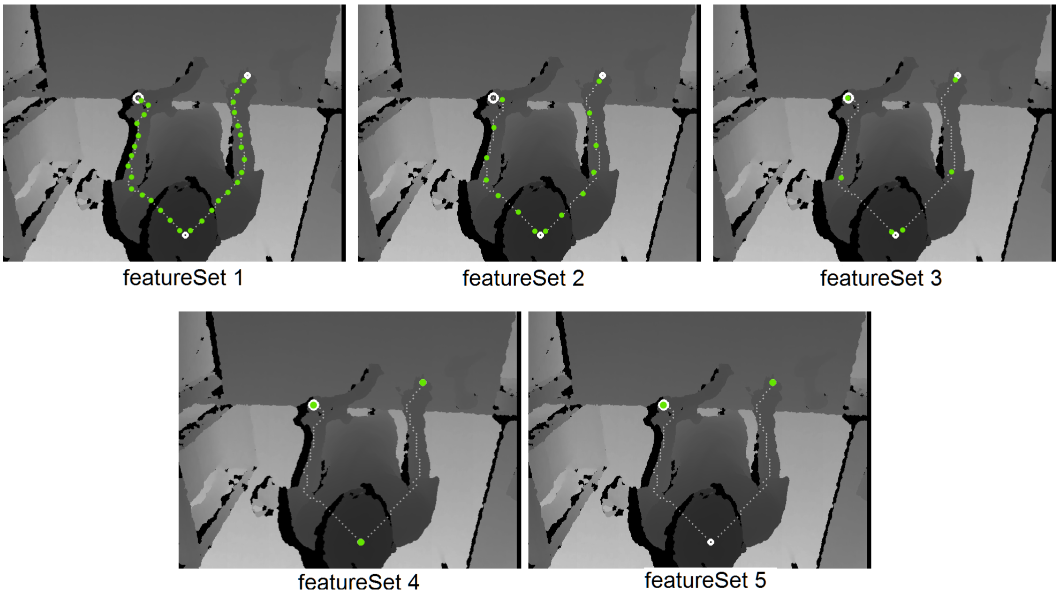
We therefore test (which we had not done in our previous works) five different sets of features, listed in Table 1 and illustrated on Figure 4. They all contain 3D locations of the two hands, completed with varying number of other upper-body posture information.

Table 1 Definitions of the five sets of features compared.

featureSet 1	15 samples of each shortest path + head and two hands 3D locations
featureSet 2	7 samples of each shortest path + head and two hands 3D locations
featureSet 3	3 samples of each shortest path + head and two hands 3D locations
featureSet 4	head and two hands 3D locations
featureSet 5	two hands 3D locations

4.1.3 Gesture classification algorithm

To classify the technical gestures performed, we have chosen to use *discrete* Hidden Markov Models (HMM). They are probabilistic models for classification of sequential discrete data. Given a continuous-valued vector of features deduced from estimated top-viewed posture (see part 4.1.2), we first need to quantize these data



featureSet 1

featureSet 2

featureSet 3

featureSet 4

featureSet 5

Fig. 4 Illustration of our five sets of features tested and compared for recognition of technical gestures (see 4.1.2 for their definition).

in order to obtain *discrete-valued* observations. For this step, we use K-Means clustering. This method aims at partitioning observations into a fixed number K of clusters. Each observation belongs to the cluster with the nearest centroïd. For our study, as already described in (Coupeté et al, 2015) and (Coupeté et al, 2016b), we partition all computed posture-estimation feature vectors into K clusters, so each cluster corresponds to an approximate top-viewed posture. After clustering, a gesture is represented as a temporal sequence of cluster IDs, corresponding to a sequence of approximate postures.

We use these quantized data to train our discrete HMMs, one HMM for each gesture class. For recognition, each feature vector extracted from depth-image is quantized by the previously learned K-Means, and the obtained labels are afterwards used as input for the discrete HMMs to determine which gesture is currently being performed. The recognized gesture is the one associated to the HMM which has the highest probability to have generated the observations. To train our HMMs, we use the Baum-Welch algorithm, and for the recognition we use the Forward algorithm. They are both implemented in the GRT[1] open library. Figure 5 illustrates our methodology.

4.1.4 Online gesture recognition in real-time

We want to *continuously* recognize gestures *in real-time* while the operator is working, performing the technical gestures one after the other naturally (i.e without any pause between successive gestures). To this end, we use a temporal sliding window of length T. Using the Forward algorithm, we compute the likelihood for each HMM to have produced the T last observations. To filter out transient errors, we finally output as recognized gesture class the one which has been the most recognized during the 10 last positions of the sliding temporal window.

To evaluate the performance of our real-time recognition system, we use the five standard metrics listed below:

$$\overline{R} = \frac{\#(gestures\ correctly\ recognized)}{\#(gestures\ performed)} \tag{1}$$

$$R_i = \frac{\#(gestures\ i\ correctly\ recognized)}{\#(gestures\ i\ performed)} \tag{2}$$

$$\overline{P} = \frac{\#(gestures\ correctly\ recognized)}{\#(gestures\ classified)} \tag{3}$$

$$P_i = \frac{\#(gestures\ i\ correctly\ recognized)}{\#(gestures\ classified\ i)} \tag{4}$$

$$F = 2\frac{\overline{P} \times \overline{R}}{\overline{P} + \overline{R}} \tag{5}$$

in which:

- $\#(gestures\ performed)$ represents the total number of gestures of all classes performed by the operators
- $\#(gestures\ i\ performed)$ represents the number of gestures of class i performed by the operators
- $\#(gestures\ correctly\ recognized)$ represents the number of gestures correctly recognized by our system.

[1] http://nickgillian.com/grt/

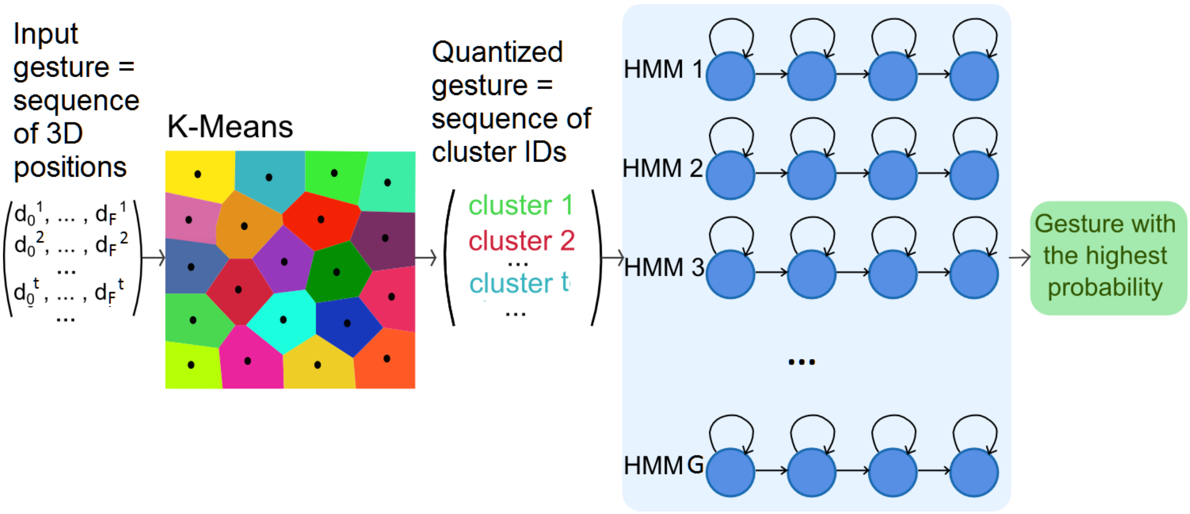

Input gesture = sequence of 3D positions
d0 1, ... , dF 1
d0 2, ... , dF 2
...
d0 t, ... , dF t
...
K-Means
Quantized gesture = sequence of cluster IDs
cluster 1
cluster 2
...
cluster t
...
HMM 1
HMM 2
HMM 3
...
HMM G
Gesture with the highest probability

Fig. 5 Gesture recognition pipeline: input gesture (left) is a temporal sequence of feature vectors of same dimension F; each continuous-valued feature vector is quantized by K-means into a *discrete-valued "approximate posture" label* (middle); the obtained temporal sequence of successive posture labels is fed one after the other into G discrete HMM (1 per gesture class); for each time-step, our system outputs the most probable current gesture class, by selecting the HMM which has current maximum likelihood.

- #(*gestures i correctly recognized*) represents the number of gestures of class i correctly recognized
- #(*gestures classified i*) represents the number of gestures classified with the label i.
- #(*gestures classified*) is equal to the sum of all the value of #(*gestes classified i*) among all the classes

The *average recall* $\overline{R}$ provides a global information on the capacity of our system to detect the gestures. The values R_i detail the separate detection ability for each class of gestures. The values P_i indicate the accuracy of our system when it outputs the corresponding gesture class ID. The *average precision* $\overline{P}$ is the average of all precisions, P_i. Finally, the F-score F is the harmonic mean of average precision and average recall, and provides a global recognition performance index.

4.2 Evaluation criteria

To evaluate our system of gesture recognition, we use two criteria. The first one, called *jackknife*, estimates the future performance of our system for a new user, from whom no gesture was used to learn K-Means and train HMMs. Our second criterion, called 80%-20%, estimates performances of our system for users of whom example gestures are included in the training set. These two criteria are illustrated on Figure 6.

4.2.1 Jackknife

Our database contains recorded technical gesture examples from N operators. To evaluate our system for an unknown user, we train it with a database composed of gesture examples from $N-1$ operators, and estimate recognition rate on gesture examples performed *only* by the last operator (not included in training set). We test all possible combinations of $N-1$ operators for training and 1 operator for recognition estimation. This evaluation criterion is illustrated on Figure 6(a).

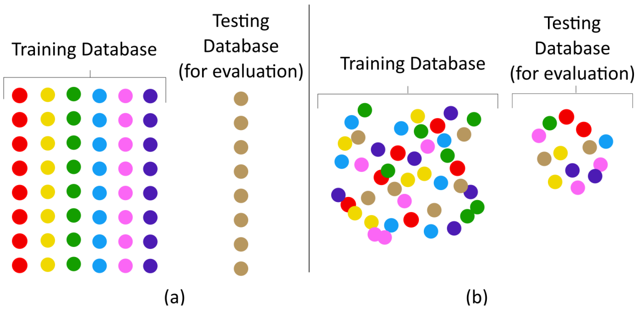

Training Database
Testing Database (for evaluation)
Training Database
Testing Database (for evaluation)
(a)
(b)

Fig. 6 Illustration of our two evaluation criteria. Each color represents an operator, and each dot a gesture example. (a): Jackknife, (b): 80%-20%

4.2.2 80% - 20%

For this evaluation criterion, we randomly divide our database in two parts. The first part is used for training and contains 80% of all gestures by all operators in our database. The second part is used for testing recognition, and is composed of the remaining 20% of our database. This evaluation criterion is illustrated on Figure 6(b). The main difference with the *jackknife* is that with the 80%-20% criterion, the system uses examples of gestures from the same operators in both training and testing databases, so it estimates recognition performance for “known” operators, i.e. included in the training set.

4.3 Use of inertial sensors *placed on tools*

To get more information on executed gestures, it can be interesting to equip the scene with additional sensors. In particular, valuable and complementary information can be obtained by *placing inertial sensors on the tools* that are used by the operator. Thus, as already reported in our previous work (Coupeté et al, 2016b), we have put an inertial sensor on the screwing-gun[2] . We use this additional data source with a “late-fusion” scheme: output of vision-based HMMs are first computed, and movement information from the tool is used only afterwards to deduce the final gesture classification result. We chose the “late-fusion” method because these data will only be used to distinguish one particular gesture class (“screwing”) against another one (“assembling”).

The screwing-gun is supposed to move only when the worker is using it to screw together two parts of motor hose. There is a conflict with the result of the HMMs classification in two cases:

- case 1: when gesture G4 (“screwing”) is recognized, while the screwing gun does **not** move
- case 2: when a gesture which is not G4 is recognized, while the screwing gun did move

For the first case, if we suppose that the inertial sensor cannot be broken, it is not possible to screw without moving the screwing-gun. Therefore, if the likelihood of the HMM for the gesture “to screw” is above a threshold, we decide that this gesture has been executed, otherwise no gesture is recognized (zero output).

For the second case, it is possible that the screwing-gun moved without being used, if the worker wants to move it from one side of the table to another for example. In this case we also look at the output likelihood of the HMM matching with the gesture “to screw”. If this likelihood is above a threshold, we replace the gesture previously recognized by “to screw”, otherwise we keep the gesture associated to the HMM with the highest likelihood.

With this method, we want to make our system more robust by correcting confusion errors that can easily occur between rather similar gesture classes.

4.4 Adaptation to a new user

As already unveiled in (Coupeté et al, 2016a), we studied the adaptation to a new user of our system. For this purpose, we experiment adaptation of the training dataset to this new user. We think that it is quite feasible in practice, and worth considering, when an operator learns to perform a new human-robot collaborative task, to record several repetitions of his gestures while he is experimenting his new task. We could have tried to apply an incremental learning algorithm: starting from the HMMs pre-trained on other users and fine-tune them with gesture examples from the new user. However, because in our case HMM training is rather fast, we decided to proceed by re-training from scratch on training database enhanced by addition of some gesture examples by the new user. As already reported in our previous work (Coupeté et al, 2016a), we also investigate the impact on recognition performance of the number of gesture examples from the new user added to original multi-users database. Figure 7 illustrates our methodology. Furthermore, in our final application, it could be possible to switch between user-specific gesture classifiers depending on the identity of current operator. We therefore perform and present here a new evaluation to compare with performance of classifiers trained only on other gesture examples of the same new user.

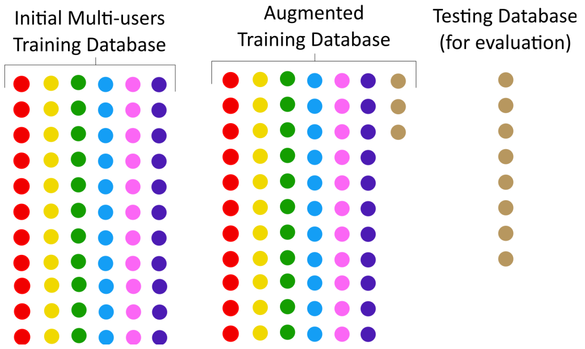

Initial Multi-users Training Database
Augmented Training Database
Testing Database (for evaluation)

Fig. 7 Our method for adaptation to a new user: the initial multi-users training set is complemented by a few gesture examples of the new user, and retrained “from scratch”. Evaluation is performed only on independant gesture examples of the same new user.

To evaluate this method, we add to the previous database a growing number of sets of gesture examples recorded from the new operator. One set is composed of one gesture of each class.

As for the *jackknife* criterion, we test all possible combinations of gesture examples from $N-1$ operators + ϵ gesture examples of the last operator to create

[2] Note that in modern factories, many tools such as screwing guns are actually connected to the assembly-line information system, so that binary information such as “moving” or “in use” can be readily available even without having to place inertial sensors on them.

the training database, and remaining gestures examples by the last operator to create the test database. Also, to avoid potential bias due to varying size of training set, we take care of maintaining a constant number of gesture examples in our databases (training and test), whatever the added number of new user gestures' sets.

5 Experiments and results

In this section, we present our results. In a first part, 5.1, we explain how we recorded the gestures to create our dataset of examples. In a second part, 5.2, we present our study to choose optimal set of features describing technical gestures. Afterwards, we provide and justify our choices of parameters (part 5.3), and then present our *online* gesture recognition results (part 5.4). Finally, we provide and analyze our gesture recognition performances after adapting our system to a new user by training set modification (part 5.5); those results are also compared with performance attainable when training our system with gestures from only the new user (part 5.5.2).

5.1 Data acquisition protocol

To have a sufficiently large dataset for testing our method, we recorded 13 "naïve" operators (among which 2 women and 11 men) aged from 25 to 60 years old (47 years old on average). Each operator has executed between 20 and 25 assembly tasks. For each assembly, between 7 and 8 successive gestures are performed by the operator. Note that operators did not have any prior knowledge or experience on the task: they were only shown how to assemble pieces together, and told that the robot would handle pieces to them, but absolutely no instruction was given to them on detailed way to execute the technical gestures; this implies that operators were actually performing the assembly task for the first time during recording, thus increasing variability even between cycles executed by same operator.

5.2 Comparison of feature sets

As explained in 4.1.2, instead of using directly as features all the body posture information that our depth-image processing algorithm provides, we try and evaluate five different sets of features (see Table 1 for their definitions and Figure 4 for their illustrations).

Table 2 shows the results obtained with these different sets of features, which is a new finding not investigated in our previous works. We can observe that best result of correct gesture recognition, 79%, are obtained when we use *only the two hands 3D locations*. When we add information which are not directly linked with the *effective* part of gestures, the recognition rates decrease. Indeed, samples from the shortest paths and head location provide information about the operator's posture, but they can vary significantly from one operator to another, and even between several executions of the same gesture by the same user. Furthermore, this finding is coherent with the results of (Chen et al, 2015) in which very good recognition rates using only hand movements as features are reported, for gestures of an operator in a set-up similar to ours.

Table 2 Rates of correct gesture recognition obtained depending on set of features. Recognition on isolated gestures, *jackknife* criterion.

15 samples head location hands locations	**7 samples head location hands locations**	**3 samples head location hands locations**	**head location hands locations**	**hands location**
65%	**70%**	**72%**	**74%**	**79%**

For the rest of presented results, we use as features only the set of two hands 3D locations.

5.3 Gesture classification algorithm parameters

As described part 4.1.3, we use a combination of K-Means and discrete HMMs to learn and recognize technical gestures. Both these algorithms have parameters that we must determine: the number of clusters for K-Means, and the number of hidden states for the HMMs. We have tested different combinations of parameters (which we had not reported in our previous publications) to determine and choose the values providing the best results. These tests are conducted on isolated gestures, i.e. gestures which are already segmented, and using jackknife criterion. Results are presented in Table 3.

We can observe that, when the number K of K-means clusters increases, the rate of correct gesture recognition gets clearly better, until K reaches 20 or 25. This was somewhat expected because more clusters implies a finer discretized description of hands postures, allowing a better distinction between different classes of gestures; and conversely when quantization is sufficiently fine, further increase of K cannot bring more improvement. As highlighted in Table 3, the recognition rate reaches a maximum of 82% of correct recognitions

Table 3 Correct gesture recognition rates as a function of number K of K-means clusters, and number S of HMM states.

		Number S of HMM states					
		5	**7**	**10**	**12**	**15**	**20**
Number K of K-means clusters	**10**	74%	75%	73%	72%	74%	73%
	15	76%	78%	78%	79%	78%	79%
	20	77%	80%	77%	78%	79%	78%
	25	76%	77%	79%	**82%**	81%	80%
	30	77%	78%	78%	80%	80%	79%

for $K = 25$ clusters and using HMMs with $S = 12$ hidden states. For the rest of this study, we use 25 clusters for K-means and 12 hidden states for HMMs as parameters of our gesture classification algorithm.

5.4 Online recognition perfomances

We need to recognize gestures *while they are performed by the operator*, and ideally even before they are finished.

Figure 8 illustrates two examples of output by our online continuous gesture recognition, each one during a complete assembly task executed by an operator. The blue line and the colors on the background represent the ground truth, i.e. the gestures which are currently performed by the operator. The red line represents the *real-time* output of our system.

On top of Figure 8, during the execution of the first gesture 1 “take a motor hose part in the robot right claw”, from 0 to 2 seconds, our system is still recognizing the previous gesture 5, “to put the assembled piece in the box”, but finally recognizes gesture 1 roughly 0.5 seconds before the end of its execution. All the following gestures are correctly recognized before they end (most of the time between 0.5 and 2 seconds in advance). During this assembly, our system wrongly recognizes gesture 3, “to join two parts of the motor hose”. During this time the operator has his two hands in front of him, while waiting for the robot to bring him the next motor piece. This posture is similar to the one observed during execution of gesture 3, this is why we can observe this mistake.

On bottom of Figure 8, one can also observe that our system correctly recognizes technical gestures performed by the operator. In this case, it can be noticed that our system sometimes outputs zero instead of a gesture class ID. This occurs when current gesture is not well-enough recognized, and our system returns a zero value, rather than risking to output a false recognition.

In the two next sections, we present our results of online gesture recognition evaluated with our two criteria, jackknife and 80%-20%.

5.4.1 Jackknife

We first evaluate with jackknife criterion the result of our system continuously recognizing gestures in real-time. The performances, depending on duration of the temporal sliding window, are presented in Table 4 for recall rates and in Table 5 for precision rates.

For recall, best results are obtained with medium duration of temporal sliding windows: 1 second and 1.5 seconds. For both window lengthes, we have a $\overline{R}$ score of 77%. With the window duration of 0.5 second the $\overline{R}$ score is lower, 65%. This window is not long enough to contain sufficient information to correctly recognize the technical gestures. For the longest sliding window, 2 seconds, we obtain a $\overline{R}$ score of 74%. With this duration, short gestures can be drowned with other information, inhibiting their correct recognition.

Table 4 Recall for *online continuous* recognition of technical gestures using data from the depth-camera and from inertial sensor on the screwing-gun. Evaluation criterion: jackknife, number of states: 12, number of clusters: 25

Length of temporal sliding window	**Recall**					
	$\overline{R}$	R_1	R_2	R_3	R_4	R_5
0.5 s	**65%**	54%	62%	38%	94%	83%
1 s	**77%**	68%	79%	60%	95%	87%
1.5 s	**77%**	67%	75%	64%	95%	84%
2 s	**74%**	65%	64%	64%	95%	82%

We observe for precision a trend similar to the one obtained for recall. Better results are obtained with longer temporal sliding windows, and the best one is obtained with a window duration of 1 second, reaching a $\overline{P}$ score of 84%.

Table 5 Precision for *online continuous* recognition of technical gestures using data from the depth-camera and from inertial sensor on the screwing-gun. Evaluation criterion: jackknife, number of states: 12, number of clusters: 25

Length of temporal sliding window	**Precision**					
	$\overline{P}$	P_1	P_2	P_3	P_4	P_5
0.5 s	**79%**	57%	81%	71%	92%	80%
1 s	**84%**	68%	87%	79%	92%	86%
1.5 s	**83%**	67%	87%	76%	92%	84%
2 s	**83%**	53%	76%	77%	98%	90%

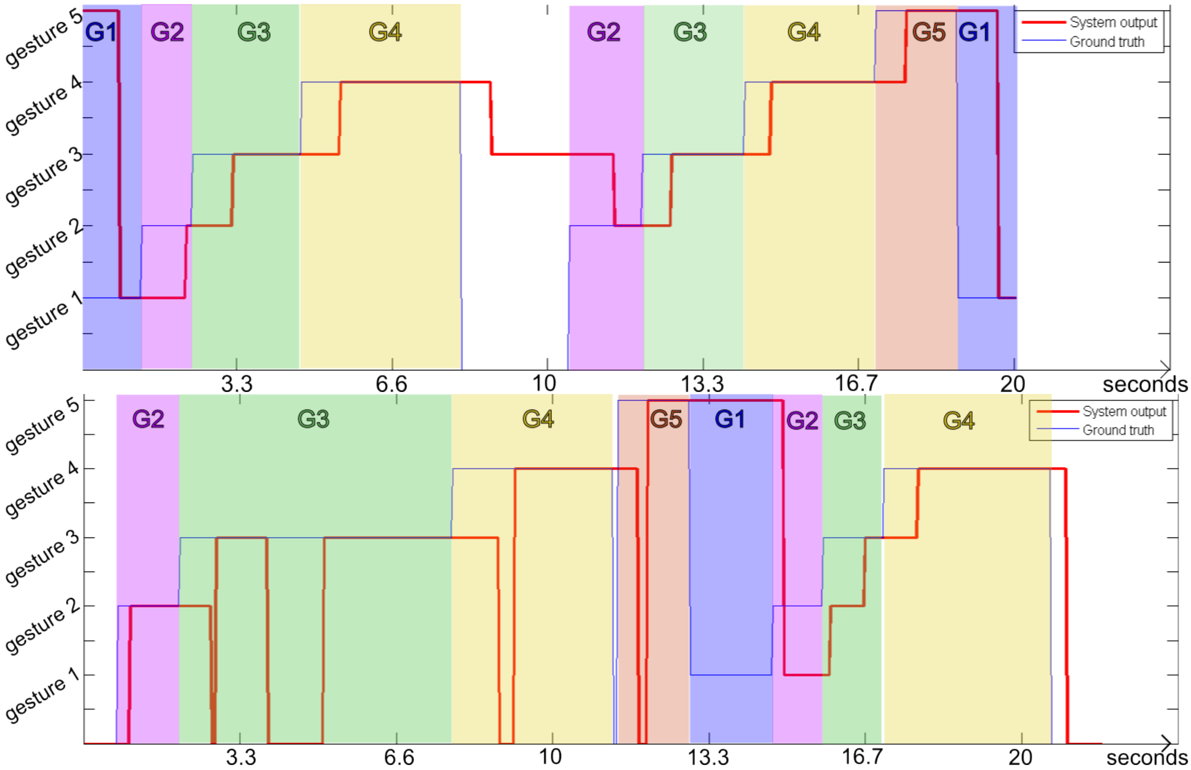


Fig. 8 Examples of online continuous gesture recognition. Blue line and colors on the background: ground truth, Red line: our system output. Note that, in order to avoid false recognitions, our system sometimes ouputs '0' (as can be seen on second example) instead of a gesture class when it is unsure about the type of currently executed gesture.

5.4.2 80% - 20%

We also evaluate our system using the 80%-20% criterion. The results are presented in Tables 6 and 7 for recall and precision respectively.

These results follow a trend similar to that observed with *jackknife* criterion. The best results of recall are obtained for medium-sized temporal sliding windows, with a duration of 1 second and 1.5 seconds. For these windows, we reach a recall of 85% and a precision of 82%. As expected, we can observe that recall is higher when using the 80%-20% criterion than with *jackknife*. Indeed, with the 80%-20% criterion, the system already *"knows"* the operator, i.e. some of *his* gesture examples were used to train the HMMs. These results motivated us to explore a method to adapt our system to a new user by modifying the training set.

5.4.3 Recognition delays

As can be seen on Figure 8 by comparing change instants of blue and red lines, our algorithm performs *early* recognition of gestures, in the sense that the operator's action is often correctly classified BEFORE the gesture is finished, and even sometimes a rather short time after it is initiated. We also have quantitatively

Table 6 Recall for *online continuous* recognition of technical gestures using data from the depth-camera and from inertial sensor on the screwing-gun. Evaluation criterion: 80%-20%, number of states: 12, number of clusters: 25

Length of temporal sliding window	**Recall**					
	$\overline{R}$	R_1	R_2	R_3	R_4	R_5
0.5 s	**80%**	38%	85%	76%	95%	77%
1 s	**85%**	44%	87%	87%	96%	80%
1.5 s	**85%**	55%	86%	80%	95%	86%
2 s	**81%**	64%	88%	80%	95%	81%

Table 7 Precision for *online continuous* recognition of technical gestures using data from the depth-camera and from inertial sensor on the screwing-gun. Evaluation criterion: 80%-20%, number of states: 12, number of clusters: 25

Length of temporal sliding window	**Precision**					
	$\overline{P}$	P_1	P_2	P_3	P_4	P_5
0.5 s	**74%**	68%	95%	55%	92%	89%
1 s	**80%**	70%	94%	75%	92%	89%
1.5 s	**82%**	70%	91%	77%	92%	89%
2 s	**73%**	67%	77%	76%	92%	89%

evaluated these delays between initiation of an action and the instant when it is correctly recognized by our system. As shown in Table 8, this delay is typically between 1 and 1.5 seconds, to be compared with gestures durations which vary between 1.5 s and 3 s; it is interesting to note that for gesture classes G3 and G4, recognition occurs on average respectively 0.9 s and 0.6 s *before* the end of the gesture.

Table 8 Time delay between gesture initiation and its recognition (averages and standard deviations, in seconds), compared to gesture average duration (in seconds)

Gesture class	mean gesture duration	*mean recognition delay*	*St.Dev.*
G1	1 s	1.1 s	0.4 s
G2	1.1 s	1.3 s	0.6 s
G3	2.5 s	1.6 s	1.2 s
G4	2 s	1.4 s	0.6 s
G5	1.7 s	1.6 s	0.6 s

5.4.4 Comparison with Dynamic Time Warping (DTW)

In order to assess if our particular recognition method (K-means+discrete-HMMs) has an important contribution to our final results, we have also conducted tests using DTW (Dynamic Time Warping) instead. As mentioned in section 2.2.3, DTW is a very commonly used technique for gesture recognition, so it provides a useful baseline result. As can be seen in Table 9 the recognition performance is much lower with DTW than with K-means+HMMs. This can be explained by the quite large intra-class variability of gesture execution in our application, because DTW models each gesture class by one single template, which makes it less suitable for our technical gestures. This hypothesis is confirmed by the fact that the drop of recognition rate for a new user (Jackknife criteria) compared to a “known” user (80%-20% criteria) is much higher with DTW (−10%) than with K-means+discrete-HMMs (−3%), which means our recognition method is clearly more robust to gestural variability.

Table 9 Comparison of gesture recognition performance between DTW and K-means+discrete-HMMs): average F-score, average Precision and average Recall

Algo (evaluation criteria)	$\overline{F}$	$\overline{P}$	$\overline{R}$
DTW (Jacknife)	**39%**	**47%**	**34%**
DTW (80%-20%)	**50%**	**59%**	**44%**
K-means+HMMs (Jacknife)	**80%**	**83%**	**77%**
K-means+HMMs (80%-20%)	**83%**	**82%**	**85%**

5.5 Adaptation of gesture recognition to a new user

5.5.1 Adaptation of training dataset

We can observe on results presented above that our rates of correct gesture recognition are better when the system “knows” the user, i.e. was trained with a dataset containing at least some gesture examples performed by him. Indeed, we obtain better results with the 80%-20% criterion than with *jackknife*. This observation motivated us to adapt the training database to a new user, as explained in part 4.4.

Our approach consists in adding one or several sets of gesture examples executed by the new operator to the training database. For the comparison to be fair, we randomly remove gestures from the original multi-operators dataset when we add gestures by the new operator, in order to maintain the same size of training and testing datasets for all tests.

The results (using a 1 second long temporal sliding window) are compared on Table 10. Both recall and precision increase to reach 89%, when 15 sets of gesture examples have been added. Table data are plotted on top-left of Figure 9, showing recall precision and F-score as a function of the number of sets added in the training base.

Table 10 Precision and recall of technical gestures for online recognition, after an adaption of the training base with an increasing added number of gesture examples from the new user. Data from the depth-camera and inertial sensor on the screwing-gun.

	Number of sets of gesture added								
	1	**2**	**3**	**4**	**5**	**7**	**10**	**12**	**15**
$\overline{P}$ (%)	84	84	86	85	86	85	86	86	**89**
$\overline{R}$ (%)	84	84	87	87	86	87	88	88	**89**
F (%)	84	84	86	86	86	88	86	87	**89**

The maximum improvement is quite large (+12% of recall and +5% of precision, compared to the initial jackknife results in Tables 4 and 5). Interestingly, it appears that even when adding only a small number (≤ 5) of gesture sets, the improvement is significant (+9% recall and +3% precision). It can also be seen on curves plotted on top-right of Figure 9 that the first 5 added sets bring significantly more improvement by set. The recall and precision continue to increase when more gesture sets are added, but the impact of each set

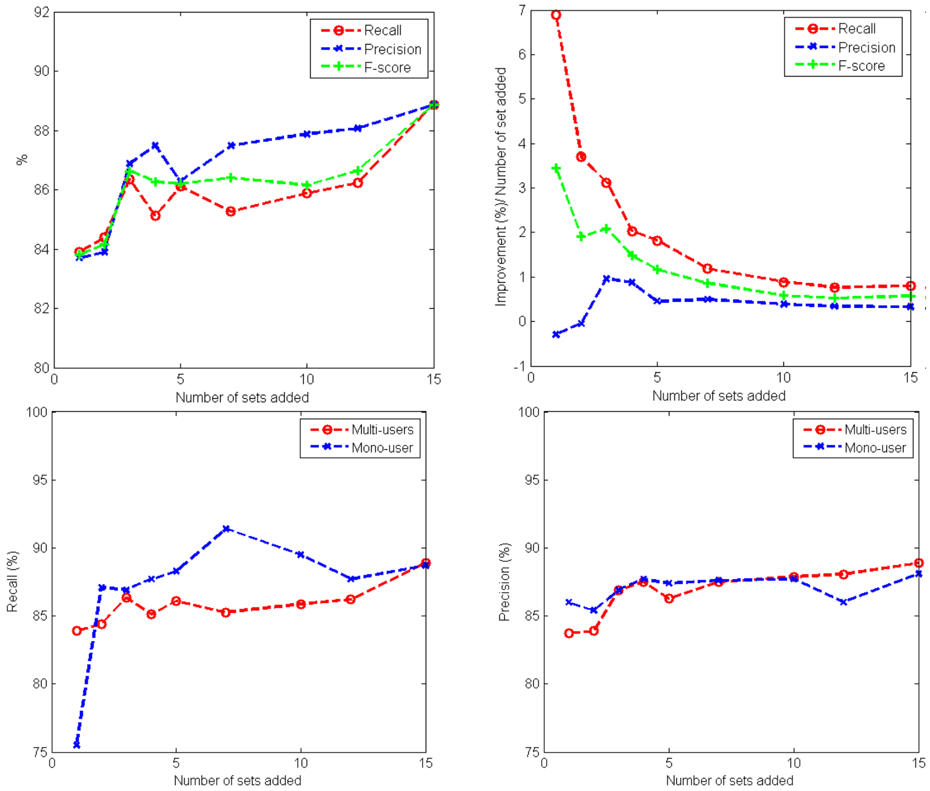


Fig. 9 Adaptation of gesture recognition to new user. On top, results obtained by training on multi-users database enriched by addition of some gesture examples from a new user (left: recall, precision and F-score as a function of the number of gesture sets added; right: improvement contribution brought by each new added set). On bottom, comparison of recall (on left) and precision (on right) between the performances attained by adding gestures example sets by new user to multi-users training database (red lines), and the result of training *with gesture examples ONLY by the new user* (blue lines).

decreases and converges around 1% of improvement for each new set added.

These observations show that adding to the training dataset a relatively small number of gesture examples from a new user can, after full retraining, significantly improves gesture recognition performances for this new user. Using operator-specific personalized gesture classifiers is therefore desirable, and easily feasible by re-training from scratch after very slight augmentation of an initial multi-users training database.

5.5.2 Comparison with training on mono-user datasets

Since we test a system adapted to a new user by modifying the training base, one can wonder what would be the performances of a system trained on gesture examples recorded *only* from this new user. We therefore also evaluate gesture recognition results, in case our system is trained and tested on databases composed only of gesture examples from the same operator. We conduct this evaluation, which is a new study compared to our previous work, for an increasing number of gestures in training set, the size of the test database being constant. Results are shown in Table 11. Not surprinsingly, recognition performances strongly increase when the number of gesture sets grows, particularly from 1 set (F-score = 80%) to 7 sets (F-score = 89%). Improvement is much slower for further addition of gesture sets.

Table 11 Precision, recall and F-score of technical gestures *online* recognition with *mono-user* training and testing databases. Data from the depth-camera and inertial sensor on the screwing-gun.

	Number of gestures sets in training								
	1	**2**	**3**	**4**	**5**	**7**	**10**	**12**	**15**
$\overline{P}$ (%)	86	85	87	88	87	88	88	86	**88**
$\overline{R}$ (%)	75	87	87	88	88	**91**	89	88	89
F (%)	80	86	87	88	88	89	89	87	88

Plots on bottom of Figure 9 compare recognition performances attained by using mono-user training dataset with those presented in 5.5.1, where a multi-users training database was enriched with a few gesture examples by the test user. For recall (graph on the bottom-left), results of the mono-user-trained system are globally better than those obtained from the multi-users adapted system; however recall rates become quite similar for higher number of gesture sets. For precision (bottom-right graph), results for the mono-user system and the multi-user adapted system are globally very close. However, *above ten sets, the precision from the system based on a multi-users adapted training base are better than those obtained with the mono-user system.*

These results of precision and recall suggest that a multi-users adapted system can be more robust than a mono-user system, which is a new finding. This could be explained by the fact that a multi-user adapted training set contains a greater quantity of ways to perform the same gesture than a same-size mono-user database. Hence the learnt HMMs are more general, allowing a better precision during recognition.

6 Conclusions and perspectives

In this study, we presented an experiment *on a real prototype* in which we continuously recognize, online and in real-time, technical gestures performed by operators on an assembly-line. This study highlights the feasibility to recognize technical gestures in such context *using only non-intrusing sensors.*

We use a depth-camera with a top-view to minimize possible occlusions on the collaborative task. We choose the gesture classes to be recognized so as to optimize coordination between the robot and operator.

We propose an algorithm for estimating upper-body posture (especially hands positions) using geodesic distances between upper-body pixels and the head's top. We highlight that features directly linked to the *effective part* of gestures (hands movements) lead to better recognition results than using user's upper-body global posture.

We show that our system can recognize technical gestures in real-time, even for users not included in training database examples. We also propose a method to adapt gesture classification to a new user, by moderate enrichment of the training set. We reach 91% recall and 88% precision during online multi-users gesture recognition.

Furthermore, we highlight that training on a dataset adapted to a new user by addition of rather few gesture examples can lead to better precision of gesture recognition than learning a totally user-specific classifier for each operator, trained with only his own example gestures. This could be due to the fact that a multi-users database includes more variability of gestures' execution, leading to more robustness.

As for perspectives, we currently work on handling parasite gestures, which can be performed by operators while they are working, but are not technical gestures. We also plan to investigate use of different lengthes of temporal sliding windows for each gesture class, to take into account their unequal average durations. It could also be interesting in a future work to analyze if there could be a relation between situation of actual hands positions 6D vector within clusters along the gesture trajectory, and success or failure of the gesture recognition by the HMMs.

Finally, since our gesture recognition methodology (choice of feature set, classification pipeline, adaptation to new user) is rather general, it could be used in application contexts other than manufacturing assembly-lines, for example in assistance and service collaborative robotics.

Acknowledgements This research benefited from the support of the Chair 'PSA Peugeot Citroën Robotics and Virtual Reality', led by MINES ParisTech and supported by PEUGEOT S.A. The partners of the Chair cannot be held accountable for the content of this paper, which engages the authors´ responsibility only.

References

Aarno D, Kragic D (2008) Motion intention recognition in robotassisted applications. Robotics and Autonomous Systems 56(8):692–705

Bannat A, Bautze T, Beetz M, Blume J, Diepold K, Ertelt C, Geiger F, Gmeiner T, Gyger T, Knoll A, Lau C, Lenz C, Ostgathe M, Reinhart G, Roesel W, Ruehr T, Schuboe A, Shea K, Stork genannt Wersborg I, Stork S, Tekouo W, Wallhoff F, Wiesbeck M, Zaeh MF (2011) Artificial Cognition in Production Systems. IEEE Transactions on Automation Science and Engineering 8(1):148–174

Biswas KK, Basu SK (2011) Gesture recognition using microsoft kinect. In: The 5th International Conference on Automation, Robotics and Applications, IEEE, pp 100–103

Bregonzio M, Gong S, Xiang T (2009) Recognising action as clouds of space-time interest points. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1948–1955

Bulling A, Blanke U, Schiele B (2014) A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys 46(3):1–33

Calinon S, Billard A (2004) Stochastic Gesture Production and Recognition Model for a Humanoid Robot. In: Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on, pp 2769–2774

Chen C, Jafari R, Kehtarnavaz N (2016) Fusion of depth, skeleton, and inertial data for human action recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 2712–2716

Chen CP, Chen YT, Lee PH, Tsai YP, Lei S (2011) Real-time hand tracking on depth images. In: 2011 Visual Communications and Image Processing (VCIP), IEEE, pp 1–4

Chen F, Zhong Q, Cannella F, Sekiyama K, Fukuda T (2015) Hand gesture modeling and recognition for human and robot interactive assembly using hidden markov models. International Journal of Advanced Robotic Systems 12(4):48

Chen L, Wei H, Ferryman J (2013) A survey of human motion analysis using depth imagery

Cherubini A, Passama R, Crosnier A, Lasnier A, Fraisse P (2016) Collaborative manufacturing with physical human–robot interaction. Robotics and Computer-Integrated Manufacturing pp 1–13

Corrales Ramón JA, García Gómez GJ, Torres Medina F, Perdereau V (2012) Cooperative tasks between humans and robots in industrial environments. InTech

Coupeté E, Manitsaris S, Moutarde F (2014) Real-time recognition of human gestures for collaborative robots on assembly-line. In: 3rd International Digital Human Modeling Symposium (DHM2014), Tokyo, Japan, p 7 p.

Coupeté E, Moutarde F, Manitsaris S (2015) Gesture Recognition Using a Depth Camera for Human Robot Collaboration on Assembly Lines. Procedia Manufacturing 3:518–525

Coupeté E, Moutarde F, Manitsaris S (2016a) A User-Adaptive Gesture Recognition System Applied to Human-Robot Collaboration in Factories. In: 3rd International Symposium On Movement and Computing (MOCO'16), Thessalonique, Greece

Coupeté E, Moutarde F, Manitsaris S, Hugues O (2016b) Recognition of Technical Gestures for Human-Robot Collaboration in Factories. In: The Ninth International Conference on Advances in Computer-Human Interactions, Venise, Italy

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numerische Mathematik 1(1):269–271

Dollar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior Recognition via Sparse Spatio-Temporal Features. In: 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, IEEE, pp 65–72

Dong L, Wu J, Chen X (2007) A Body Activity Tracking System using Wearable Accelerometers. 2007 IEEE International Conference on Multimedia and Expo pp 1011–1014

Dragan AD, Bauman S, Forlizzi J, Srinivasa SS (2015) Effects of Robot Motion on Human-Robot Collaboration. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15, ACM Press, New York, New York, USA, pp 51–58

Hägele M, Schaaf W, Helms E (2002) Robot Assistants at Manual Workplaces: Effective Co-operation and Safety Aspects. Proceedings of the 33rd ISR (International Symposium on Robotics) 7-11

Hamester D, Jirak D, Wermter S (2013) Improved estimation of hand postures using depth images. In: 2013 16th International Conference on Advanced Robotics (ICAR), pp 1–6

Hoffman G, Breazeal C (2007) Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In: Proceeding of the ACM/IEEE international conference on Human-robot interaction - HRI '07, ACM Press, New York, New York, USA, p 1

Joo SI, Weon SH, Choi HI (2014) Real-time depth-based hand detection and tracking. The Scientific World Journal

Junker H, Amft O, Lukowicz P, Tröster G (2008) Gesture spotting with body-worn inertial sensors to detect user activities. Pattern Recognition 41(6):2010–2024

Ke Y, Sukthankar R, Hebert M (2007) Spatio-temporal Shape and Flow Correlation for Action Recognition. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1–8

Laptev I, Lindeberg T (2003) Space-time interest points. Proceedings Ninth IEEE International Conference on Computer Vision 1:432–439

Lenz C, Nair S, Rickert M, Knoll A, Rosel W, Gast J, Bannat A, Wallhoff F (2008) Joint-action for humans and industrial robots for assembly tasks. In: RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication, IEEE, pp 130–135

Liu J, Zhong L, Wickramasuriya J, Vasudevan V (2009) uwave: Accelerometer-based personalized gesture recognition and its applications. Pervasive and Mobile Computing 5(6):657 – 675

Luo J, Wang W, Qi H (2013) Group Sparsity and Geometry Constrained Dictionary Learning for Action Recognition from Depth Maps. In: The IEEE International Conference on Computer Vision (ICCV), pp 1809–1816

Migniot C, Ababsa F (2013) 3d human tracking from depth cue in a buying behavior analysis context. In: 15th International Conference on Computer Analysis of Images and Patterns (CAIP 2013), pp 482–489

Oikonomopoulos A, Patras I, Pantic M (2005) Spatiotemporal salient points for visual recognition of human actions. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics) 36(3):710–719

Reyes M, Domínguez G, Escalera S (2011) Featureweighting in dynamic timewarping for gesture recognition in depth data. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp 1182–1188

Rickert M, Foster ME, Giuliani M, By T, Panin G, Knoll A (2007) Integrating Language, Vision and Action for Human Robot Dialog Systems. In: Universal Access in Human-Computer Interaction. Ambient Interaction, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 987–995

Schrempf OC, Hanebeck UD, Schmid AJ, Worn H (2005) A novel approach to proactive human-robot cooperation. In: ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005., IEEE, pp 555–560

Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., IEEE, vol 3, pp 32–36 Vol.3

Schwarz LA, Mkhitaryan A, Mateus D, Navab N (2012) Human skeleton tracking from depth data using geodesic distances and optical flow. Image and Vision Computing 30(3):217–226

Sempena S, Maulidevi NU, Aryan PR (2011) Human action recognition using dynamic time warping. In: Electrical Engineering and Informatics (ICEEI), 2011 International Conference on, IEEE, pp 1–5

Shi J, Jimmerson G, Pearson T, Menassa R (2012) Levels of human and robot collaboration for automotive manufacturing. In: Proceedings of the Workshop on Performance Metrics for Intelligent Systems - PerMIS ’12, ACM Press, New York, New York, USA, p 95

Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Cook M, Moore R (2011) Real-time Human Pose Recognition in Parts from Single Depth Images. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA, CVPR ’11, pp 1297–1304

Wang H, Ullah MM, Klaser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. In: Procedings of the British Machine Vision Conference 2009, British Machine Vision Association, pp 124.1–124.11

Wang P, Li W, Gao Z, Zhang J, Tang C, Ogunbona PO (2016) Action Recognition From Depth Maps Using Deep Convolutional Neural Networks. IEEE Transactions on Human-Machine Systems 46(4):498–509

Xia L, Chen CC, Aggarwal JK (2012) View invariant human action recognition using histograms of 3D joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 20–27

Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hidden Markov model. In: Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Comput. Soc. Press, pp 379–385

Zhang H, Parker LE (2011) 4-Dimensional Local Spatio-Temporal Features for Human Activity Recognition. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp 2044—-2049

Zhu HM, Pun CM (2012) Real-time Hand Gesture Recognition from Depth Image Sequences. 2012 Ninth International Conference on Computer Graphics, Imaging and Visualization pp 49–52

Capture, modeling and recognition of expert technical gestures in wheel-throwing art of pottery.

S. MANITSARIS[1, 2, 3],
A. GLUSHKOVA[1, 2, 4],
F. BEVILACQUA[3],
AND
F. MOUTARDE[2]
University of Thessaly[1] – Greece, MINES ParisTech[2] - France,
IRCAM[3] - France, University of Macedonia[4] - Greece

This research has been conducted in the context of the ArtiMuse project that aims at the modeling and renewal of rare gestural knowledge and skills involved in the traditional craftsmanship and more precisely in the art of the wheel-throwing pottery. These knowledge and skills constitute the Intangible Cultural Heritage and refer to the fruit of diverse expertise founded and propagated over the centuries thanks to the ingeniousness of the gesture and the creativity of the human spirit. Nowadays, this expertise is very often threatened with disappearance because of the difficulty to resist to globalization and the fact that most of those "expertise holders" are not easily accessible due to geographical or other constraints. In this paper, a methodological framework for capturing and modeling gestural knowledge and skills in wheel-throwing pottery is proposed. It is based on capturing gestures using wireless inertial sensors and statistical modeling. In particular, we used a system that allows for online alignment of gestures using a modified Hidden Markov Model. This methodology is implemented into a Human-Computer Interface, which permits both the modeling and recognition of expert technical gestures. This system could be used to assist in the learning of these gestures by giving continuous feedback in real-time by measuring the difference between expert and learner gestures. The system has been tested and evaluated on different potters with a rare expertise, which is strongly related to their local identity.

General Terms: Technical gestures, Know-how, Modeling, Recognition, Wheel-throwing

Additional Key Words and Phrases: Perception, HCI, Inertial sensors, Machine-Learning

1. INTRODUCTION

Cultural expression is not limited in architecture, monuments, collection of objects, or music. It also includes fragile intangible live expressions, which involve knowledge and skills. Such characteristics can be human gestures, the color of our voice or facial expressions. They are controlled by the intelligence of the human creativeness depicted in music, dance, singing, theatre, human skills and handicraft. This kind of culture has been termed Intangible Cultural Heritage (ICH), which is manifested inter alia in oral traditions and expressions, performing arts (*music, dance, theatre etc.*) social practices, knowledge and practices concerning nature, universe and the traditional craftsmanship [Unesco 2003]. In traditional craftsmanship, the rare gestural knowledge primarily consists of hand and finger motions.

Authors' addresses: [1, 2, 3]S. Manitsaris E-mail: sotiris.manitsaris@uom.edu.gr. [1, 2, 4]A. Glushkova, E-mail: alina.glushkova@uom.edu.gr. [3]F. Bevilacqua, E-mail: frederic.bevilacqua@ircam.fr. [2]F. Moutarde, E-mail: fabien.moutarde@mines-paristech.fr. [1]Rural Space Lab, Department of Planning and Regional Development, University of Thessaly, Volos, Greece. [2]Robotics Lab, MINES ParisTech, Paris, France. [3]Sciences and

Technology for Music and Sound Lab, Ircam-CNRS-Université Pierre et Marie Curie,Paris, France. [4]Multimedia Technology and Computer Graphics Lab, Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece.

© 2001 ACM 1530-0226/07/0900-ART9 $5.00 DOI 10.1145/1290002.1290003 http://doi.acm.org/10.1145/1290002.1290003

The project “ArtiMuse” aims at proposing a multidisciplinary research approach for the gesture recognition methodologies applied in musical and handicraft interactions. In order to preserve these gestural skills that require high expertise it is necessary to identify, record, analyze, model and recognize them. This paper presents the development of a methodology for the modeling of kinematic aspects of technical gestures using gesture recognition technologies based on wireless inertial sensors. This methodology has been implemented, applied and evaluated on the gestural skills of two expert potters for simple objects, such as bowls. The Human-Computer Interface called “ArtOrasis” has been developed for gesture capturing, modeling and recognition of expert gestures. It also proposes the time alignment of a gesture compared to a model gesture, which is a quite promising perspective for a performance comparison between expert and learner.

2. STATE OF THE ART

The study of human gestures has been of special interest in different research fields. In the ICH domain in particular, body and hand gestures are important means of communication, of expression and of creativity. Preservation and transmission of handicraft skills is often done by studying and analyzing recordings, verbal descriptions and documents. However, the important role of gestures has lead several researchers to model them using motion capture and gesture recognition technologies.

2.1 E-DOCUMENTATION AND DIGITAL ETHNOGRAPHY

Most existing methods for skill preservation are based on verbal descriptions of movements, often in conjunction with multimedia content such as graphic / photographic material [D. Chevallier, 1991]. Another widespread method is the video recording of skills and techniques of the gesture accompanied with verbal commenting. This method was applied to preserve knowledge of technical gestures in power stations in France during the manipulation of different technical tools. A video camera has been placed on the helmet of the workers to record their movements [Le Bellu 2012].

In the field of ICH, for many years, ethnologists have studied the characteristics (*arts and techniques, oral traditions and living expressions*) of groups and communities in their surroundings. [D. Chevallier, 1991]. Enora Gandon worked on wheel-throwing gestures in a cultural context. She studied the impact of the cultural background of the human on the development of motor skills in Wheel-throwing art [E. Gandon, 2011]. Kuo-An Wang studied the case of weaving Chinese traditional items with Bamboo [Kuo-An Wang et. al, 2011]. He created a digital archive with approximately1200 objects accompanied by images and videos presenting the gestures involved in the creation of the objects. Through meetings and interviews conducted with the craftsman, Kuo-An Wang has identified a set of 20 basic gestural patterns of weaving. Then, he connected each of the digitized objects file with a combination of those patterns.

However some significant limitations can be identified in these methods. While describing a gesture on a piece of paper using photos or figures, the gesture is limited in two dimensions and it does not represent any realistic information about how the gesture has been performed. In the case of video recording, the gesture is also represented in two dimensions but still limits the information that can be extracted.

2.2 MOTION CAPTURE AND GESTURE RECOGNITION

The use of innovative technologies for motion capture permit to overcome some of the limitations mentioned above, to achieve a faithful record of the gesture and model it stochastically. A significant number of studies have been based on various techniques for modeling and recognition of gestures based on motion capture. These technologies can be subdivided in 3 categories: a) marker-based, b) marker-less and c) inertial motion sensors.

Marker-based approaches use optical-markers and active computer vision, which require expansive commercial systems, such as Vicon Peak or Optitrack. This type of sensor has been used for the modeling of music performances of a violin player [Rasamimanana N. et al. 2009, Demoucron M. et al., 1994]. One of the most important limitations of the marker-based gesture recognition systems is that they are not robust to occlusions.

Marker-less technologies do not require subjects to wear special equipment for tracking and are usually based on passive computer vision approaches. For example, Microsoft Kinect is a low-cost depth camera that provides good results for the recognition of global body postures, such as dance gestures [Raptis, 2011]. It provides Cartesian representation of the human motion and it is usually used for tracking joints of the body. Nevertheless, it is less precise for hand gestures. To solve this problem there is an ongoing research aiming at the creation of a hand skeletal model for finger detection. A hand skeletal model for depth images, provided by the PMD CamBoard Nano time of flight camera, has been applied to capture music-like finger gestures [Dapogny et al., 2013] based on Shotton's algorithm and pixel wise classification through Random Decision Forest. This model is currently being developed and adapted for pottery-like finger gestures, as a training database is required. An elaborated algorithm is also necessary for scene and object segmentation in the case of technical gesture recognition in wheel-throwing art of pottery. Moreover, the depth cameras are self and scene occlusion-dependent and the development of hand/finger skeletal model for the capturing of finger gestures in pottery is a challenging approach in a medium term vision.

Inertial Motion Sensors [R. Aylward et al., 2006, T. Coduys et al., 2004, T. Todoroff, 2011, E. Fléty et al. 2011] or commercial interfaces, such as the Wii joystick [D. Grunberg, 2008], permit to track gesture features continuously and in real-time. These sensors have been tested and used in dance and music performances [Bevilacqua et al., 2007, 2010]. For example, inertial sensors have been used for motion capture aiming at the archaeological reconstruction, understanding and interpretation of different possibilities of use of an ancient Iron Age roundhouse [S. Dunn et al, 2011].

2.3 PRESSURE MEASUREMENT AND MECHANICAL STRESS

In case of wheel throwing pottery, a study has been conducted to evaluate potters' skills by taking into consideration the mechanical characteristics of the objects created by the potters. The Von Mises stress index [J. Lemaitre et al., 2004] has been used to measure the mechanical stress operating in the object and it has been related to the throwing difficulty, proposing thus an idea for potters' skills assessment [E. Gandon et al., 2011]. In this object-oriented approach, the pressure is measured by analyzing the mechanical characteristics of the created object after completion, and not in real-time during the creation of the vessel. Therefore, despite its interest, this mechanical stress estimation cannot easily be used to build an interactive pedagogical tool, because interactivity requires real-time analysis of the gesture performed. Since designing this type of tool is one of our final goals, our approach for the preservation and modeling of gestural skills is based on the analysis of the kinematic aspects of the gesture required for the creation of the object, which can be easily captured and analyzed in real-time.

3. METHODOLOGY

In the methodology proposed below, the goal is not a simple video recording with verbal descriptions, nor just the digitalization of information. The objective is to study and to model rare gestural know-how involved in handicrafts, to gather data about different biomechanical, kinematic aspects of a technical gesture (*distances between the hands, angles of the vertebral axis, rotations of the joints, gesture's trajectory etc.*), as well as the body postures and to create information about its various parameters. The modeling of the gestural know-how and the effective recognition of gestures have been done in different methodological phases as described below and represented in the figure 1.

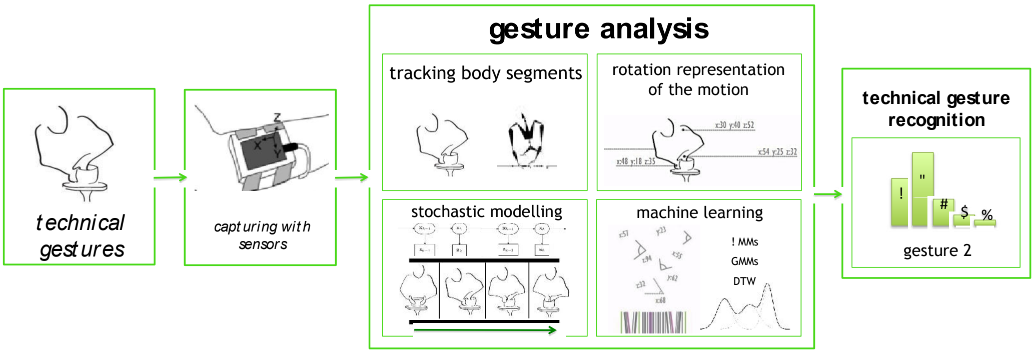


Fig. 1. Gesture recognition pipeline based on the Animazoo suit of wireless motion sensors for the upper-part of the potter's body.

Capturing, modeling, online recognition and time alignment, are fundamental stages and extensions towards the creation of an interactive pedagogical tool for the transmission of gestural skills based on sensorimotor learning. This comparison could be done in real time in order to provide a sonic or optical feedback to the learner, and thus drive him/her to correct his/her gestural errors.

3.1 HANDICRAFT AND EXPERT SELECTION

The first step is to select the type of handicraft and the craftsman to be studied. The handicraft selected should respect some criteria and be compatible with the goal of the research and with its' technical conditions and constraints. Since we are using gesture capture technologies, one of the important criteria for technical feasibility and application of the methodology is to avoid use of specific tools during gesture execution.

3.2 IDENTIFICATION OF THE EFFECTIVE GESTURES

The second phase concerns the identification of effective gestures. Effective gestures have a direct impact on the material and it is important to identify them and to distinguish them from other auxiliary gestures. This goal can be achieved in interaction and collaboration with the expert, by conducting interviews and observing him while working. He should show and describe a complete sequence of gestures, effective and auxiliary, in order to define a dictionary $GD = \{G_i\}_{i\in\mathcal{N}}$ of his/her effective gestures.

3.3 GESTURE CAPTURING

In the context of expert technical gestures, a possible simplification of the complexity of the human body can be based on rotation of segments, which is the case of the data stream (*observation vector*) provided by the sensors. Therefore, the expert's body segments are tracked and the sensor captures kinematic properties about angle rotations in 3D space and record them as a sequence of observation vectors $Y_{0:k} = \{y_1 \cdots y_k\}_{k\in\mathcal{N}}$, where $y_{j_{j\in\mathcal{N}}}$ is one observation vector and k the total number of observations. The vector $y_j = [y_1^j \cdots y_n^j]_{j,n\in\mathcal{N}}$ represents the n kinematic descriptors for a given time stamp j.

3.4 GESTURE MODELING AND MACHINE-LEARNING

3.4.1 GESTURE REPRESENTATION

Once the data acquisition is completed, then $Y_{0:k}$ should be normalized. Euler angles, Quaternions are possible rotation representations of the motion. Euler rotation is a rotation about a single Cartesian axis. According to the Euler's Rotation Theorem, every orientation can be described as a rotation from some other reference orientation as a sequence of three elemental rotations (*precession, nutation, and intrinsic rotation*). Quaternions are representing orientations and rotations of objects in 4D, where there is one real axis and three imaginary axes (*i, j, and k*). Another way to model the rotations of the motion of the human body is the Direction Cosine Matrix (DCM), which is based on a triad of unit vectors. The rotation is described by specifying the coordinates of the triad of unit vectors in its current position, based on a non-rotated coordinate axes that is used as a reference.

3.4.2 GESTURE MODELING AND ONLINE CHARACTERISATION

One of the difficulties in gesture recognition is that the same gesture can be performed in a variety of ways, in particular the change in speed of execution. For this reason, one of the authors has developed a system called "Gesture Follower" that can be seen as a hybrid approach between Dynamic Time Warping (DTW) and Hidden Markov Models (HMM) [Bevilacqua 2007, 2010]. This system is a template-based method, which allows the use of a single gesture to define a gesture class. This requirement is necessary due to the limited access of gesture data in our application. Nevertheless, we use a HMM formalism to compute in real-time computation measures between the template and the incoming data flow.

As described by Rabiner (1989), Hidden Markov Model can be used to model recorded time series (*training procedure*) and to compute the likelihoods (*one per HMM*) that the hidden state sequence $X_{0:k} = \{x_1 \dots x_k\}_{k\in\mathbb{N}}$, generated the new observation sequence $Y_{0:k} = \left\{y_1 \dots y_k\right\}_{k\in\mathbb{N}}$. In our case, we are interested in the following information:

- The likelihood $\psi(x,y)$ that the observation sequence was produced by the different models. The sequence with the maximum likelihood ψ_{max} to generate $Y_{0:k}$ indicates the gesture G_i from the gesture dictionary $GD = \{G_i\}_{i\in\mathbb{N}}$.
- The likeliest state sequence $X_{0:k}$ of the observation sequence $Y_{0:k}$. This state-sequence allows for obtaining a time-warping between the model and the observed sequence.

In order to greatly simplify the learning procedure and to guarantee a high temporal precision in the gesture modeling, we associate each template to a state sequence. We define one state for each sample data of the template (*applying a constant sampling rate*). We compute the different likelihoods in real time using the well-known for-ward procedure that return the results incrementally, and that can be implemented efficiently even in the case of models with a large number of states [Bevilacqua 2010].

4. CASE STUDY

4.1 SELECTION OF THE WHEEL-THROWING POTTERY AND OF THE POTTERS

For the implementation of this methodology the wheel-throwing pottery has been chosen mainly for two reasons. First of all, because of the high social and cultural value of this traditional profession for the local communities of the Macedonia Region in northern Greece (*community of the potter A*) and of the French Riviera (*Côte d'Azur*) in Southern France, (*community of the potter B*) which is also famous for the large number of ceramists working there. An important number of associations and independent experts are actively promoting this handicraft in these regions. A list of candidate craftsmen has been established. Two experts from the regions above have been selected. They detain a very high level of expertise in wheel-throwing pottery art and also important pedagogical experience. The potter A is teaching this handicraft in a centre for therapy and social reintegration of people with substance dependencies and the potter B is a senior craftsman with more than twenty years of experience in practicing wheel-throwing pottery. The second criterion is more technical and linked to our system's technical characteristics. The wheel-throwing pottery is based on gesture control of the material. There should be no interference of the hands of the potter and his material with specific tools or other intermediate mechanisms.

4.2 BASIC GESTURES FOR THE CREATION OF A BOWL

The two selected expert potters presented to us a complete sequence of gestures that are used to create a bowl with a simple shape (*a bowl*) with different quantities of clay.

It has been asked to the potter A to create 5 bowls of 18-20 cm of diameter, 10 cm of height, with approximately 1.3 kg of clay. Additionally, it has been asked from the potter B to create bigger bowls of the same shape of those created by potter A, with 20-23 cm of diameter and 13 cm of height with 1.75 kg of clay.

After meticulous observation of video recordings of the gestures of the two potters and after the interviews conducted with them, we have concluded on the following 4 basic gestural phases of creating a simple bowl (Figure 3).

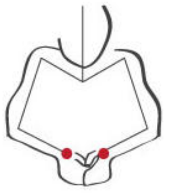
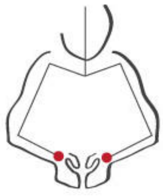
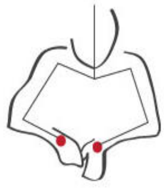
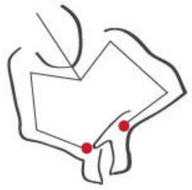
<table>
<tr><td></td><td colspan="2"></td><td></td><td></td><td colspan="2"></td></tr>
<tr><td>4 basic gestural phases</td><td colspan="2"><b>P_1
Centering and bottom opening</b></td><td><b>P_2
The raise</b></td><td><b>P_3
The first configuration</b></td><td colspan="2"><b>P_4
The final configuration and removing</b></td></tr>
<tr><td>Potter A 4 gestures</td><td colspan="2">G_1^A
Centering and bottom opening</td><td>G_2^A
The raise</td><td>G_3^A
The first configuration</td><td colspan="2">G_4^A
The final configuration and removing</td></tr>
<tr><td>Potter B 6 gestures</td><td>G_1^B
Centering the clay</td><td>G_2^B
Opening the bottom</td><td>G_3^B
The raise</td><td>G_4^B
The first configuration</td><td>G_5^B
The final configuration</td><td>G_6^B
Removing the object</td></tr>
</table>

Fig 3. Basic phases and gestures per potter for the creation of a bowl

Since the dimensions of the bowls created by the two experts are different, the required gestures inside the gestural phases are also different. Obviously, there is a common track between the gestures of the two experts, even if the dimensions of their objects are slightly different. The bigger an object is, the more gestural work it requires. Consequently, since the object of the potter A is smaller, the 4 basic gestures that we have identified correspond exactly to the 4 basic gestural phases presented above. For the bigger object, the potter B has more clay to manage and he is paying more attention to shape refining.

They are also considered as representative of the high-level wheel-throwing pottery skills, since the creation of the object takes in case of the potter A only 60-75 sec and for potter B 120-140 sec. The fluidity and the speed of potters' gestures as well as the hand coordination are elements that constitute the basis of the rare know-how and gestural skills.

All gestures have duration of 15-25 seconds. The centering and bottom opening P_1 consists of fixing of the clay on the wheel, hands are pressing steadily on the material aiming at the opening of the bottom. Then, the potter's hands are picking up the clay, defining the height of the bowl through the second gesture, P_2 the raise of the clay.
Then the body posture is changing, slightly turning on the right or on the left side for the first configuration of the shape, for P_3. Precise finger gestures are specifying the basic form of the object. The fingers of the one hand are fixing the clay and of the other are forming the object. His hands are too close to each other, touching the inner and outer sides of the clay respectively. After this stage, the potter is making the final configuration of the shape P_4. His fingers are controlling and equalizing the bowl thickness and at the end the potter passes a very fine wire between the bowl and the wheel in order to take the bowl.

4.3 EXPERIMENTATIONS: CAPTURING THE POTTER'S GESTURES

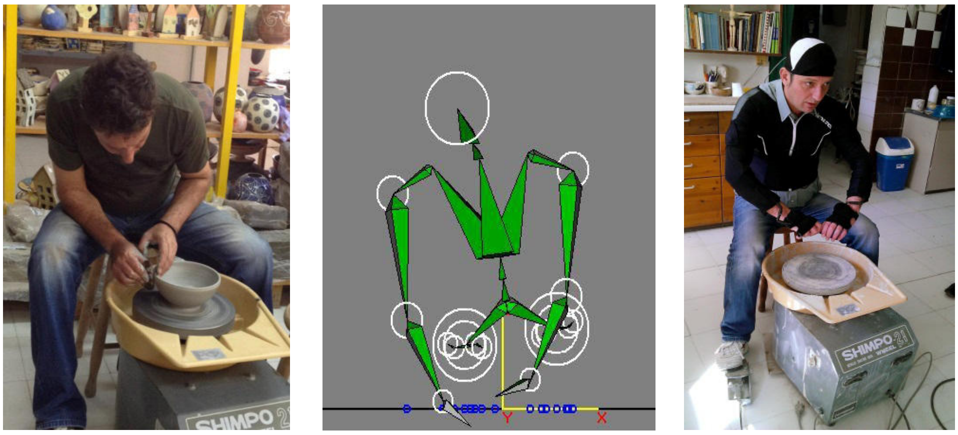
After the definition of the above effective gesture the potters are asked to put the inertial motion capture suit that can easily provide real-time access to motion information and permits the data acquisition (Figure 4).

This suit contains 11 inertial sensors (*gyroscopes and magnetometers*) and it is covering for the upper part of potter's body, his wrists, his neck and his head. It provides an automatic filtering for the correction of magnetic disruption. It has been selected for the gesture capturing and the implementation of the methodology described below. It is occlusion-independent and it provides a high precision rotational representation of body segments.

The 11 sensors are integrated in the suit and after the calibration they provide and capture information related to the XYZ axis rotations with the use of integrated gyroscopes, accelerometers and magnetometers. The posture of the upper-part of the potter's body can be derived by the data obtained from the suit but it is not the case with his position in 3D space. These data are recorded following a hierarchical structure and more precisely the Bounding Volume Hierarchy (BVH).

Since magnetometers are used among other sensors in the suit with the inertial sensors, the quality of data captured can be influenced by magnetic disturbances. During the first day of data acquisition with the potter B these disturbances were very strong since he was using an old model of wheel, containing many metallic devices. Despite the fact that data are online corrected by the system if weak magnetic disturbances are identified, the data acquired at the first day were of a very bad quality. For this reason another data acquisition session has been realized with the use of a more modern wheel with less magnetic disturbances.

The following table I, lists the different parts of potter's body, which motions have been captured for the preservation of his gestural know-how. Some of them may play a more important role in the technical gesture depending on the type of handicraft, but all the following body articulations are involved in the performance of the gesture of the craftsman. We are also aware about the important role of fingers in wheel-throwing process. Finger tracking constitutes an important step and we are currently working on the creation of a skeletal model trained on the potter's finger gestures.



(a) (b) (c)

Fig 4. (a) The potter A, Theodoros Galigalidis, pottery teacher at the "Therapy Center for Depending Individuals" in Thessaloniki, is performing the G_3 (*first configuration*). (b) Skeleton reconstruction. (c) The potter is wearing the suit with the inertial sensors for preliminary analysis.

Table I. Body segments and gesture descriptors for rotations in 3D space

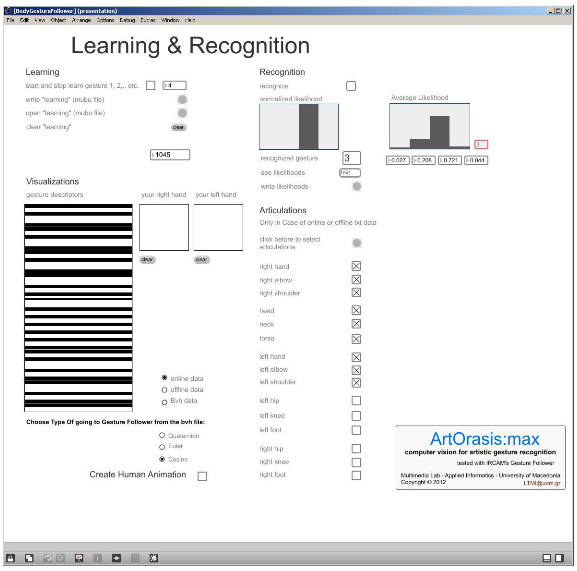
Body segments	Gesture descriptors (y_n^k)
- Spine	
- Right Shoulder	
- Right Arm	- Direction Cosine Matrix, y_{99}^k
- Right Forearm	
- Right Palm	
- Left Shoulder	- Euler angles, y_{33}^k
- Left Arm	
- Left Forearm	- Quaternions, y_{44}^k
- Left Palm	
- Hips	
- Head	

4.4 POTTER'S GESTURE MODELLING WITH THE ARTORASIS SYSTEM

After the gesture capturing using the inertial sensors, the data is normalized in [-1, 1] using Euler angles, Quaternions and Directors Cosine Matrix as described in the methodology.

A prerequisite for the creation and the application of our methodology was to design the ArtOrasis system and interface (Figure 5). This gesture recognition system is entirely implemented in MaxMSP, an interactive programming environment that uses the Jitter toolbox and it aims at the recognition of technical gestures. ArtOrasis can also be used for capturing, modeling, and recognition. It also provides functionalities for the visualization of the skeleton of the craftsman.

The machine learning engine of the ArtOrasis is based on a hybrid Hidden Markov Model and Dynamic Time Warping approach, which is implemented into the Gesture Follower (GF) [Bevilacqua et Al. 2007] patch for MaxMSP (developed by the IMTR research team of IRCAM).


[BodyGestureFollower] (presentation)
File Edit View Object Arrange Options Debug Extras Window Help
clear
clear
clear
0.027
0.208
0.721
0.044
3
text

Learning & Recognition

Learning
start and stop learn gesture 1, 2,.. etc. □ 4
write "learning" (mubu file) ●
open "learning" (mubu file) ●
clear "learning"
1045

Recognition
recognize □
normalized likelihood
recognized gesture 3
see likelihoods
write likelihoods ●

Average Likelihood

Visualizations
gesture descriptors
your right hand
your left hand

● online data
○ offline data
○ Bvh data

Choose Type Of going to Gesture Follower from the bvh file:
○ Quaternion
○ Euler
● Cosine

Create Human Animation □

Articulations
Only in Case of online or offline txt data
click before to select articulations ●
right hand ☒
right elbow ☒
right shoulder ☒
head ☒
neck ☒
torso ☒
left hand ☒
left elbow ☒
left shoulder ☒
left hip □
left knee □
left foot □
right hip □
right knee □
right foot □

ArtOrasis:max
computer vision for artistic gesture recognition
tested with IRCAM's Gesture Follower
Multimedia Lab - Applied Informatics - University of Macedonia
Copyright © 2012
LTMI@uom.gr

Fig 5. Screenshot of the interface of the ArtOrasis system presenting the learning and recognition phases

In case of wheel throwing pottery 11 segments of the human skeleton listed in the table below have been selected and used for the training of ArtOrasis system. Concerning the different gestures separation, the training has been based on the 4 effective gestures identified during the second stage.

According to the model defined previously the user (*researcher, potter, learner*) of ArtOrasis can define and choose which are the most important parts of the body that participate in the execution of a technical gesture and train the system based on ones. This stage corresponds to the machine learning phase of the methodology. The training of the gesture recognition system is also based on the effective gesture separation defined in second methodological step. After the training of the system the last step is the gesture recognition that is evaluated below.

5. EVALUATION

One of the final goals of our research is the design of a real-time pedagogical tool that can help transmission, and thus preservation of gestural know-how. To attain this goal, we need to compare gesture realization by an apprentice with the recorded and modeled gestures by experts. A pre-requisite before estimating such similarity, is the automated recognition by the system of what particular step the apprentice is trying to perform. For this reason we are convinced that online technical gesture recognition is essential for the comparison of handicraft skills between apprentices and expert. Furthermore, the segmentation of the data captured into a set of specific gestures, and the training of models, provides the data with a semantic dimension.

In order to validate our approach, and evaluate the recognition accuracy of the system for all the P_i, it has been asked to each of the expert potters to create five bowls. All the gestures G_i^A from the potter A and G_i^B from the potter B that are involved in all the four phases P_i have been recorded in real conditions (*co-articulated gestures and without rest*). It has to be mentioned that very often, expert craftsmen are not available to create many copies of exactly the same object since this procedure is considered as a creative art process or because of ageing. Nevertheless, in case of the potter A the repeatability of his gestures can be considered as being of a high level, since he was very concentrated and careful in the way he performed the gestures. In case of the potter B the repeatability is of a medium level since he is easily disturbed in his everyday work by external elements (neighbors visiting his atelier, etc.)

The gesture recognition rates have been evaluated based on the « jackknife » method [Abdi, Williams 2010]. In our case, jackknifing means estimation of the recognition accuracies for manually segmented gestures (*isolated gestures*) by using subsets of the available gestural data. The basic idea behind the jackknife variance estimator lies in systematically recomputing the statistic estimate leaving out one or more observations at a time from the sample set.

Practically, a dataset contains observations of all the G_i^A and G_i^B. In total, five observations for each gesture have been recorded and distinct databases for learning and test have been used in five iterations. For each iteration, one dataset is left out to be used as the learning database and train one model M_i per gesture G_i until all the data sets are used once and the four remaining datasets are used as a database for testing. Two metrics have been used to evaluate the system:

$$Precision = \frac{\#\ of\ True\_Recognitions}{\#\ of\ True\_Recognitions\ +\#\ of\ False\_Recognitions} \quad (1)$$

$$Recall = \frac{\#\ of\ True\_Recognitions}{\#\ of\ True\_Recognitions\ +\#\ of\ Missed\_Recognitions} \quad (2)$$

So, for the potter A, the first evaluation phase has been done using Euler angles after normalization for training the HMMs M_i^A. For each of the eleven body segments, one Euler angle per axis has been computed. The table II shows the results of the five iterations of the jackknifing as well as the Precision and Recall per gesture of the potter A. Twelve queries for recognition per G_i^A have been asked to the M_i^A. Both Precision and Recall were at 100%.

Table II. Precision and Recall per gesture from the potter A based on five iterations of jackknifing using Euler angles.

		Maximum likelihoods (ψ_{max})				
		M_1^A	M_2^A	M_3^A	M_4^A	**Recall**
Observations ($Y_{0:k}^A$)	G_1^A	**20**	-	-	-	100%
	G_2^A	-	**20**	-	-	100%
	G_3^A	-	-	**20**	-	100%
	G_4^A	-	-	-	**20**	100%
Precision		100%	100%	100%	100%	

The Quaternions and the Direction Cosine Matrix have also been used for the training of the $\mathrm{M_i}$ models. For both Quaternions and Direction Cosine Matrix, all the observations $Y_{0:k}$ from G_2^A and G_3^A that have been given as query to ArtOrasis gave true recognized (Recall) but there are some cases where M_2^A and M_3^A gave maximum likelihood for false recognitions (Precision). In table III we can see that all three ways of the motion representation give excellent results for the recognition of all the effective gestures.

This first experimental case shows that, at least for the creation of simple objects in wheel throwing-art, online gesture recognition based on machine-learning can be successfully applied, and can therefore be used as a first step for "capturing of gestural skills" related to pottery.

Table III. Comparative table for Precision and Recall of Euler, Quaternions and Direction Cosine Matrix (DCM) representations

	Euler		**Quaternions**		**DCM**	
	Precision	Recall	Precision	Recall	Precision	Recall
G_1^A	100%	100%	80%	94%	100%	85%
G_2^A	100%	100%	100%	90%	87%	100%
G_3^A	100%	100%	100%	87%	95%	100%
G_4^A	100%	100%	85%	100%	100%	95%
Total	**100%**	**100%**	**91%**	**92%**	**96%**	**95%**

More precisely, the Direction Cosine Matrix needs the most computational power since a 3x3 matrix has to be calculated and sent as an input to the HMMs, which may be a very important constraint for real-time applications. Taking into consideration the fact that only the upper part of the potter's body contributes in a direct way to the creation of the object while he/she is seating on a chair in front of the wheel, we can conclude that the degrees of freedom of his/her body are really reduced. Additionally, DCM are widely used on animation but not for analysis, recognition or modeling of rotations. Quaternions impose the non-Euclidean space, which cannot be easily interpreted by humans and consequently they cannot be meaningful for learning purposes.

With regards to the potter B, we have also evaluated the recognition accuracy based on jackknife method. In the table IV, the precision and recall for his G_i^B are presented. During this test we use Euler angles since they have been previously identified as the most relevant descriptor. Like in the previous example 20 queries for recognition per $\mathrm{G_i}$ have been asked to the M_i^B. The precision and recall are perfect for G_1^B to G_4^B. For G_5^B, there is one false recognition since G_4^B and G_5^B are very similar.

The difference between G_4^B and G_5^B is that, in G_5^B the potter B defines the shape with a tool and a sponge but in G_4^B he defines the shape without any tool, just with his hands. Gesture $\mathrm{G_6}$ has the lowest recognition rate because the potter was very disturbed. The repeatability of this gesture is low and it has a direct impact on its recognition rate. Even if the number of G_i^B is increased compared to G_i^A, the Precision and Recall are still excellent.

Table IV. Precision and Recall per gesture based on 5 iterations of jackknifing using Euler angles-Potter B.

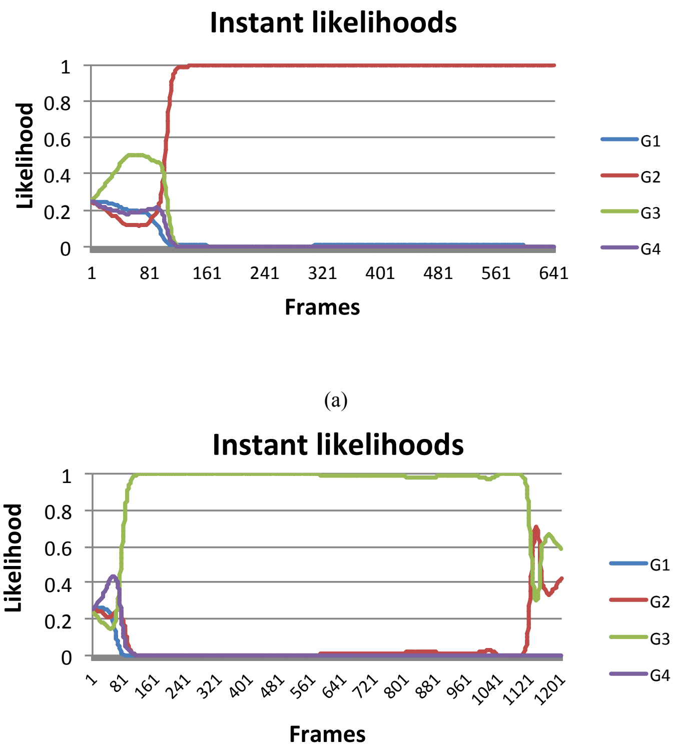
		Maximum likelihoods (ψ_{max})						
		M_1^B	M_2^B	M_3^B	M_4^B	M_5^B	M_6^B	**Recall**
Observations $Y_{0:k}^B$)	G_1^B	**20**	-	-	-	-	-	100%
	G_2^B	-	**20**	-	-	-	-	100%
	G_3^B	-	-	**20**	-	-	-	100%
	G_4^B	-	-	-	**20**	-	-	100%
	G_5^B	-	-	-	1	**19**	-	95%
	G_6^B	2	-	-	-	-	**18**	90%
Precision		91%	100%	100%	100%	95%	90%	

The possibility to do cross-potters evaluation has been rejected since it is not meaningful to compare discrete gestural skills of a rare expertise, which are strongly related with the local identity of each expert.

The recognition rate is very high but is effective after a latency of 80-120 frames (*1.5 to 2 seconds*). This latency corresponds to the time that is needed to separate the different HMM models based on their associated maximum likelihood. Co-articulation has an impact to the computation of the instant likelihood and this effect usually becomes more important during the transitions between gestures. It lasts short time periods and the gestures are not clearly recognized during these periods. More precisely, co-articulation can be defined as the fusion of distinct actions into larger and holistically perceived chunks [Hardcastle et al., 1999].

On the left of figure 6 for example, G_2^A is given to ArtOrasis. At the beginning, there is a maximum likelihood alternation effect between the M_i^A. After the first 1.5 seconds,

M_2^A has the maximum likelihood until the end of the observations. As we can see, the likelihood of M_2^A is very high comparing to other models and this can be justified by the fact that the expert repeats the gestures in a very precise way.



(a)

(b)

Fig 6. Instant likelihoods per frame using Euler angles. (a) On the left G_2^A and (b) on the right G_3^A are given as test inputs to the models.

On the first figure a) of the figure 6, there is still a latency of about 1.5 second before M_3^A becomes the maximum likelihood. Then, this effect is normalized and almost instantly the M_3 gives the maximum likelihood for about 1040 frames. Just after the frame 1121, the maximum likelihood starts to alternate between M_2^A and M_3^A due to co-articulation. In G_2^A (*the raise*), the potter needs to clean his hand before to take a small tool and starts the first configuration of the shape. During the co-articulation phase, the right hand goes under the left and vice-versa, which is a common phase with the gesture G_2^A also.

In parallel, we applied a Levene's test in order to detect the equality of variances between the Euler angles for the potter A. According to this test, the variances of the four gestures in three axis are not equal. By applying the One-Way-ANOVA test, we observe that the mean values for all the four gestures are not equal on the axis X by comparing pairs of gestures. Also, equalities are extracted between G_1^A - G_2^A on the axis Z and G_2^A - G_3^A on the axis Y. The conclusion of the statistical analysis is that the four gestures are different in terms of means and this fact contributes to the very high recognition rates. The results of the One-Way-ANOVA for the potter B are very similar to those of the potter A.

As it has been discussed before, the recognition is useful for the system in order to distinguish the gestures between them. But when the gesture is correctly recognized then it is important to have information about its temporal evolution (*time index*) compared to its model. This can be used not only to measure the different performances of the same potter, but also as a way to measure the distance between the performances between expert and learner in real-time. To do this, a temporal rescaling or time warping of the gesture can be directly obtained from our hybrid DTW-HMM approach. Time alignment experiments have been done for the right palms of the two potters. In figure 7, the time alignment of G_1^B for the potter B is shown.

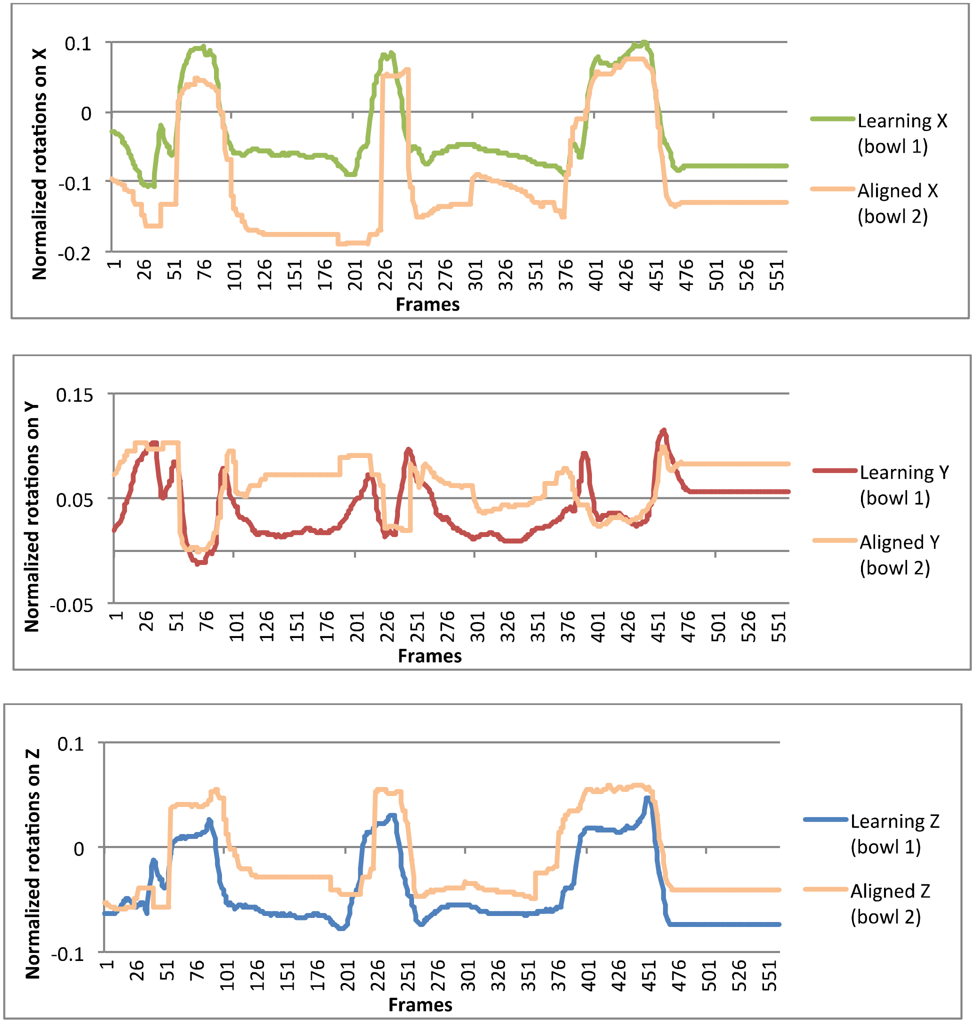


Fig.7 Example of aligned data from the right palm of the potter B for G_1^B (*Opening the Bottom*)

Normalized rotations of the first bowl for the right palm have been given to ArtOrasis as a learning sequence. Then, the normalized rotations of the second bowl have been given both as recognition and time alignment sequence. The results presented above show that for the G_1^B the aligned sequence fits well with the one that has been used for the learning of the HMMs for all three axis. Finally, from this alignment, it is possible to examine precisely where the main differences occur between the different sequences. . As previously mentioned, such difference could be displayed in real-time and offered thus feedback to the potter.

CONCLUSION

Conscious of the need for transmission of the ICH and for its preservation, we proposed a methodology for wheel-throwing gestural know-how preservation through capturing, modeling, online recognition and time alignment of different performances for the most fundamental gestures. In order to validate this methodology, the technological prototype ArtOrasis has been developed. It is able to capture rotations of body segments using inertial sensors and recognize the expert postures and gestures based on machine learning techniques. Both methodology and system have been evaluated on basic expert gestures for the creation of a bowl with the help of different gesture descriptors. Through experiments conducted on two different case studies of rare expertise, we show that gesture recognition with machine-learning can be successfully applied to the creation of simple objects in wheel-throwing pottery. This illustrates that technical gesture recognition can be used for modeling of gestural skills, which is a first required step for “capturing of ICH” in a form facilitating its transmission by interactive pedagogical tools. The high level of recall and precision of gestures recognition is justified by the fact that there are no equalities of variance but also by the expertise of the potters, who repeated the gestures in a very precise way.

The long-term goal of this research is the development of the appropriate methodology and technology for collecting, recording, classifying and modeling of hand gestures that constitute a rare know-how in various types of handicrafts. Since the role of finger touching on the material is very important, a future goal of this research is to extend the current methodology by combining it with computer vision in order to capture finger movements as well. To propose a completed methodology it would be also interesting to combine kinematic data (upper part of the body and fingers) with kinetic information about the pressure brought on the clay. But also the object detection and scene segmentation would give precious information about the evolution of the creation process and also about the progression of the gesture. The results of this study will not only contribute to the preservation of this gestural know-how but also to the development of a system aiming at the transmission. These results could be also used for renewal of ICH by proposing gestural metaphors for the creation of augmented musical performances based on handicraft gestures.

ACKNOWLEDGEMENTS

The research project is implemented within the framework of the Action «Supporting Postdoctoral Researchers» of the Operational Program "Education and Lifelong Learning" (Action’s Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the Greek State.

REFERENCES

Aylward, R., Daniel Lovell S., and Joseph A. Paradiso, 2006, A Compact, Wireless, Wearable Sensor Network for Interactive Dance Ensembles, In Proceedings of *BSN 2006, The IEEE International Workshop on Wearable and Implantable Body Sensor Networks*, Cambridge, Massachusetts, April 3-5, pp. 65-70

Abdi, H., Williams, L.J., 2010, « Jackknife », In Neil Salkind (Ed.), Encyclopedia of Research Design. Thousand Oaks, CA: Sage.

Bakis, R. 1976. Continuous speech recognition via centisecond acoustic states. In Proc. 91st Meeting of the Acoustic Society in America .

Bevilacqua, F., Guédy, F., Schnell, N., Fléty E. and Leroy N., 2007, ‘Wireless sensor interface and gesture-follower for music pedagogy’. In Proceedings of *the International Conference of New Interfaces for Musical Expression*, New York, USA, pp 124-129.

Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F. and Rasamimanana, N., 2010 ‘Continuous realtime gesture following and recognition’, *LNAI 5934*, pp.73–84.

Chevallier D., 1991, « *Savoir faire et pouvoir transmettre: Transmission et apprentissage des savoir faire et des techniques* », Edition MSH France

Coduys T., Henry C. and Cont A., 2004, TOASTER and KROONDE: High-Resolution and High-Speed Real-time Sensor Interfaces, In Proceedings of the *International Conference on New Interfaces for Musical Expression*, Hamamatsu, Japan.

Dapogny A., de Charette R., Manitsaris S., Moutarde F., Glushkova A., 2013, Towards a Hand Skeletal Model for Depth Images Applied to Capture Music-like Finger Gestures, *10th Int. Symposium on Computer Music Multidisciplinary Research (CMMR'2013),* Marseille France

Demoucron M., Askenfelt A. and Caussé R., 1994, Observations on bow changes in violin performance. In Proceedings of *Acoustics, Journal of the Acoustical Society of America*, volume 123, page 3123.

Dunn S., Woolford K., Barker L. et al, 2011, Motion in Place: a Case Study of Archaeological Reconstruction Using Motion Capture, CAA2011 - Revive the Past: In Proceedings of the *39th Conference in Computer , Applications and Quantitative Methods in Archaeology*, Beijing, China. 12-16 April

Dymarski P., 2011, *Hidden Markov Models, Theory and applications*, Published by InTech Croatia

Fléty E., Maestracci C., 2011, Latency improvement in sensor wireless transmission using ieee 802.15.4, Proceedings of the International Conference on New Interfaces for Musical Expression, pages 409–412, Oslo, Norway

Gandon E., 2011, *Influence of cultural constraints in the organization of the human movement: proposition of a theoretical framework and empirical support through the example of pottery-throwing* (France / India Prajapati / India Multani Khumar), PhD thesis, Marseille University, France

Gandon, E., Casanova, R., Sainton, P, Coyle, T., Roux, V., Bril, B., & Bootsma, R.J., 2011, A proxy of potters’ throwing skill: ceramic vessels considered in terms of mechanical stress., *Journal of Archaeological Science, 38, 1080-1089.*

Grunberg D., 2008 Gesture Recognition for Conducting Computer Music. Retrieved January 10, from: http://schubert.ece.drexel.edu/research/gestureRecognition

Hardcastle, W., and Hewlett, N., 1999 *Coarticulation: theory, data and techniques*. Cambridge: Cambridge University Press.

Le Bellu S., Le Blanc B., 2010, How to Characterize Professional Gestures to Operate Tacit Know-How Transfer? *The Electronic Journal of Knowledge Management* Volume 10 Issue 2 (pp142-153)

Lemaitre, J., Chaboche, J.L., 2004, Mécanique des Matériaux Solides. Dunod, Paris.

Raptis M., Kirovski D., Hoppe H., 2011, Real-Time Classification of Dance Gestures from Skeleton Animation. Eurographics/ In Proceedings of *ACM SIGGRAPH Symposium on Computer Animation.*

RABINER, L. R., 1989, A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Rasamimanana N. and Bevilacqua F., 2009, Effort-based analysis of bowing movements : evidence of anticipation effects. *The Journal of New Music Research*, 37(4): 339 – 351

Todoroff T., 2011, Wireless digital/analog sensors for music and dance performances. In Proc. the *International Conference on New Interfaces for Musical Expression*, pages 515–518, Oslo, Norway

Wang K.A., Liao Y.C. , Chu W.W. , Yi-Wu Chiang J., Chen Y.F. and Chan P.C., 2011, Digitization and Value-Add Application of Bamboo Weaving Artifacts, C. Xing, F. Crestani, and A. Rauber (Eds.): ICADL 2011, *LNCS 7008*, pp. 16–25, 2011. Springer-Verlag Berlin Heidelberg

UNESCO 2003. 1998. *Segmentation and Categorization of Phonemes in Continuous Speech*. Technical Report TR-CST25JUL98, Center for Sensor Technology, University of Pennsylvania.

Received: 12 April 2016 | Revised: 26 April 2017 | Accepted: 16 July 2017
DOI: 10.1111/jcal.12210

WILEY Journal of Computer Assisted Learning

ORIGINAL ARTICLE

Gesture recognition and sensorimotor learning-by-doing of motor skills in manual professions: A case study in the wheel-throwing art of pottery

Alina Glushkova[1] | Sotiris Manitsaris[2]

[1] Multimedia, Security and Networking Laboratory, University of Macedonia, Greece

[2] Centre for Robotics, MINES ParisTech, PSL Research University, France

Correspondence
Alina Glushkova, Multimedia, Security and Networking Laboratory, University of Macedonia, Greece.
Email: alina.glushkova@uom.edu.gr

Funding information
European Social Fund, Grant/Award Number: FP7; Intangible Treasures—Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures, Grant/Award Number: FP7-ICT-2011-9-600676-iTreasures; i-Treasures project; European Commission

Abstract

This paper presents a methodological framework for the use of gesture recognition technologies in the learning/mastery of the gestural skills required in wheel-throwing pottery. In the case of self-instruction or training, learners face difficulties due to the absence of the teacher/expert and the consequent lack of guidance. Motion capture technologies, machine learning, and gesture recognition may provide a way of overcoming such issues. The proposed methodology is used to record and model expert gestures and then to compare this model in real time with the gestures performed by the learner. Differences in kinematic aspects such as hand distances are detected, and optical/sonic sensorimotor feedback is provided to the learner by the system, alerting him/her when errors occur and guiding him/her to achieve better results. In the case described here, the system was evaluated with 11 learners. With the use of our system, the gestural performance of learners during self-training has been improved in comparison to cases of self-training without computer assistance.

KEYWORDS

augmented multimedia feedback, embodiment, gesture modelling and recognition, interactive learning and training, motor skills, sensorimotor learning

1 | INTRODUCTION

Manual professions are not limited to a set of tools or special devices; they also demand experience that involves gestural know-how that is controlled by hand movement intelligence and human creativity. Such experience is not limited only to the implementation of simple manual tasks but coexists with tacit knowledge. This tacit knowledge and the necessary motor skills are diffused within communities of practice, which interact with each other and ensure that the necessary knowledge and skills are transmitted from one generation to the next. Motor skills may include specific body postures to make preparatory adjustments as well as the manipulation of tools, objects, or materials used by the hands and fingers. Gestural know-how is considered at one and the same time to be traditional, contemporary, and living because it not only refers to past knowledge but also to contemporary experience that evolves through time. This transmission of experience and knowledge has an economic, social, and cultural value, because it affects not only the well- being of the job holder or artisan but also the sustainable development of society in general.

Within this context, the transmission of gestural know-how is typically made person-to-person by establishing a mirroring system between the expert and the learner. Using this “me-to-you” observation system (“me” as expert and “you” as learner), the learner perceives and understands the movements of the expert as meaningful action and not simply as displacement of space. Nevertheless, access to this expertise of gestural skills as professional know-how can sometimes be difficult to achieve due to geographical or time constraints or due to the limited number of experts that hold these skills. Within industry especially, the availability of experts for teaching or training purposes may be limited due to lack of time, in the sense that extracting an expert from his/her everyday workspace may impact on productivity. Also, in relation to craftsmanship, a number of craft traditions may include “secrets” that have been transmitted not only from one generation to another but also to a limited number of people within a particular community.

Gestural know-how transmission remains incomplete if there is no “learning-by-doing” symbiosis between the learner and his/her objects, tools, or instruments. According to the learning continuum presented by E. Dale (1969), people memorize 90% of what they are doing, 70% of what they are writing/saying, 50% of what they are watching and hearing, and 10% of what they are reading. The learning-by-doing is thus the most efficient method to acquire precise know-how. It is also important the learner to become the actor and not the observer of the situation. When learning gestures, embodied cognition requires a dynamic environment that includes interaction and manipulation (Malamed, http://theelearningcoach.com). Recent technological advances permit to imply and motivate learners, to give them this active role through serious games, simulation systems, and interactive environments (Rieber, 1996).

Depending on the type of profession, whether artistic or technical, gestures can be characterized either by expressiveness and variability or repeatability and precision. In both cases, this expertise is obtained after a long period of accumulated experience and extensive practice of gestures, whereas the expert plays an essential role in motor skills transmission. Thus, “in-person” transmission and sensorimotor learning-by-doing are closely linked notions that cannot really be dissociated. Consequently, a challenge that needs addressing is the achievement of the necessary in-person transmission and learning-by-doing, even when the expert is not present.

This paper aims to present a methodological approach and a technological paradigm for enhancing autonomy in the learning of the kinematic elements of motor skills through self-training. Motion capture, machine learning, gesture recognition technologies (GRT), and augmented multimedia feedback constitute the key elements in creating the digital metaphor of in-person transmission. In this paper, we call GRT the motion capture technologies and machine learning algorithms and their use to recognize time series of gestural data. The proof of concept is based on a series of experiments made in the wheel-throwing art of pottery.

2 | LITERATURE REVIEW

During the last century, psychologists and physiologists have studied the processes of motor skills learning and transfer. According to Newell (Newell, 1991), they are generally distinguished from perceptual, cognitive, communicative, and other skills categories. They rely on sensorimotor intelligence permitting the motor control, coordination, and action. This embodied intelligence is acquired not through cognitive processes but through senses, experiences, and consequent ambient reactions (Piaget, 1976). New mappings between motor and sensory variables are created (Wolpert, Diedrichsen, & Flanagan, 2011). When we try to obtain new motor skills and kinematics, we must also be able to link this learning to appropriate contextual cues such as objects, tasks, or environments. Roger Schank (Schank, 1997), a cognitive psychologist and an artificial intelligence theorist explains that learning from failure completes the sensorimotor learning process and permits to achieve this mapping. When a person receives feedback from his errored performance in the end, his error can also lead him to the expected correct result. The learner will thus continue trying by committing errors until the feedback conducts him to the correct performance. This idea is used for the development of the learning system presented below.

Before the transmission takes place, it is important to identify knowledge perimeters, their characteristics, and the knowledge object to be transmitted. This stage corresponds with the extraction of know-how. For many years, anthropologists, ethnologists, and experts in manual/traditional professions have been creating multimedia content, such as photographs, videos, and audio recordings in order to be able to study and transmit know-how (Chevallier, 1991). Researchers have studied expert gestures by analysing their parameters, such as trajectory and acceleration through videos and visual representations (Bril, 2011). In China, a digital archive has been created, presenting a traditional method of weaving with bamboo (Wang et al., 2011). Similar studies have been conducted in order to preserve rare dancing techniques (Kim, 2011). In the project, ChoreoSave, an important number of dances were video-recorded and the choreography was disseminated into smaller elements such as steps and figures. The preservation of gestural skills has also been studied in a more industrial context, in an electric energy production plant, where a video camera was placed on the helmet of a worker to record his gestures. These videos were used to create training material (Le Bellu & Le Blanc, 2010). Nevertheless, preserving and transmitting know-how through the use of multimedia content has considerable limitations due to the fact that (a) recordings of gestural information that are two-dimensional—because recording the video reduces the movement to two dimensions and contains limited information about the physical parameters and (b) pedagogical multimedia materials do not permit the learner to interact with them, there is little in the way of evaluation and guidance for the learner to be able to adjust movements and eliminate his/her errors.

Motion capture technology, on the other hand, overcomes some of these limitations because it provides precise information about the biomechanical aspects of a gesture such as the positioning of body joints and their rotation. At the same time, machine learning and GRT may be used to offer a more interactive learning experience. For example, in the artistic fields, a violin player's performance was captured with marker-based computer vision (Rasamimanana & Bevilacqua, 2009), an expensive and occlusion-dependent technology. In another instance, a much more accessible depth camera, Kinect, was used to capture joint positions while performing dance movements, by Raptis, Kirovski, and Hoppe (2011). Moreover, segment rotations may be captured by inertial sensors, which generally provide occlusion-independent, robust data. Sensors such as the MotionPod IGS-180 of Movea, or Animazoo Synertial, are suitable because they cover all the body segments with 18 sensors, or the upper body with 11 sensors. This type of sensor was used to capture, model and recognize expert gestures in wheel-throwing pottery (Manitsaris et al., 2014); however, inertial sensors can be conceived as invasive, and the quality of data they provide may be influenced by magnetic disturbances.

The examples mentioned above refer mostly to the use of new technologies for the preservation of know-how. They have also been used to assist learning in the arts or even in sports and medical training. In the i-Maestro project, which uses the system proposed, the violin player's movements are analysed and provide instructive optical feedback to help him/her to improve his technique (Ng et al., 2007). In speed skating on ice, sonic feedback is given to the skater, assisting him in the correction of a regular error observed in his performance (Godbout & Boyd, 2010). In the medical field, the BirthSim simulation system has been proposed (Herzig, Moreau, & Redarce, 2014). Its goal is to help obstetricians and midwives to train and improve their skills during childbirth delivery. Systems based on sensorimotor feedback have also been developed for rehabilitation purposes, using optical indications (Jégo, Paljic, & Fuchs, 2013) or continuous auditory feedback, for learning in interactive systems (Boyer, 2015). In most existing studies where feedback is used for learning purposes, the reference gesture to be learned consists of simple trajectories or is characterized by periodicity. The feedback thus provided is based on the tracking of body joints. In the present case, the reference gestures have precise kinematic features that are modelled with the use of machine learning techniques and are compared in real time with the learner's actual gestures.

3 | RESEARCH PURPOSES AND QUESTIONS

The purpose of the research was to test a novel and highly interactive embodied pedagogical system as a means to support self-training and transmit gestural know-how. Its specific aims were to study whether sensorimotor feedback can have a positive influence on the process of learning motor skills. In order to achieve this goal and to provide scientific evidence about our stated aims, three research hypotheses are proposed:

> **Hypothesis 1.** GRT contributes to the capturing and modelling of kinematic aspects of expert technical gestures.

In order to confirm this hypothesis, a number of experiments were conducted in order to test whether the machine is able to recognize expert gestures with high accuracy when a number of gestural data sets are provided. After the extraction of the know-how, it is necessary to study the relations that are established between the expert and his/her learner during the in-person transmission, as well as the difficulties the learner faces during self-training.

> **Hypothesis 2.** GRT can contribute to the evaluation of the learner's gestural performance during self-training without any sensorimotor feedback.

To confirm or refute this hypothesis, a number of metrics that evaluate learning progress are defined, these being based on both the spatial and temporal properties of the gestures. The experiments were performed after in-person transmission and without the presence of the expert.

> **Hypothesis 3.** The sensorimotor feedback has a positive impact on self-training.

To test this hypothesis, the metrics from H2 were used to evaluate whether the gesture performances improved when sensorimotor feedback was given/offered.

4 | METHODOLOGY

The methodology used is based on five stages (Figure 1), which are as follows:

1. *defining the scenario and capturing the gestures*;
2. *gesture analysis and modelling*;
3. *study of the in-person transmission and self-training*;
4. *recognition and comparison*; and, finally,
5. *sensorimotor feedback*.

The first two steps cover the extraction of (*expert*) know-how and permit the creation of the gesture models. Firstly, the aim was to define the gesture vocabulary in collaboration with the expert, in order to be able afterwards to record kinematic aspects of gestures using motion capture technology. Then, the data were preprocessed and normalized in order to be ready for use as input for our machine learning approach, which is based on hidden Markov models (HMM) and dynamic time warping (DTW; Bevilacqua et al., 2009). The main instruments for this step were semiconductive interviews for the definition of the scenario and motion capture sensors for data acquisition.

In the second methodological step, the data preprocessed must be modelled and analysed. As explained in Manitsaris, Glushkova, Bevilacqua, and Moutarde (2014) and Glushkova and Manitsaris (2015), for each gesture, a single sample was used to define a gesture class. HMMs calculated in real-time computation measures between the models and the incoming data and defined the likelihood that the hidden model would generate the incoming observation sequence. The ability of the machine to recognize different gestures would confirm or refute the first hypothesis. This ability was evaluated using the Jackknife method, and precision and recall metrics were also calculated. The modeled gestures constituted points of reference for the learner because she or he would be asked to “perform” them. In order for the machine to be able to continuously understand whether the learner is performing the gesture correctly or not, one intra-expert tolerance per gesture should be calculated. Reasonable tolerance/flexibility of the expert gesture should also be taken into consideration when the models are used for human learning. The exact value of this tolerance is calculated by using several expert repetitions of the same gesture. The goal here is to evaluate expert's degree of repeatability and to define the threshold, beyond which some parameters of the gesture are considered as incorrect. For this, we use the average or maximum of the standard deviation of 5 repetitions and use the result as an additional tolerance threshold for the learner's executions. The instruments used during this methodological steps were machine learning techniques, such as HMM and DTW, and statistical analysis notions such as standard deviation calculation.

The third step corresponds to (a) the study of expert–learner relation during the in-person transmission (*expert–learner*), (b) definition of the evaluation mechanism, and (c) the evaluation of the learner's performance after the in-person transmission and during the self-training, without any assistance (*learner alone*). Step 3a provides input for the design of the sensorimotor feedback because it must be based on the natural interaction of the learner with his or her environment

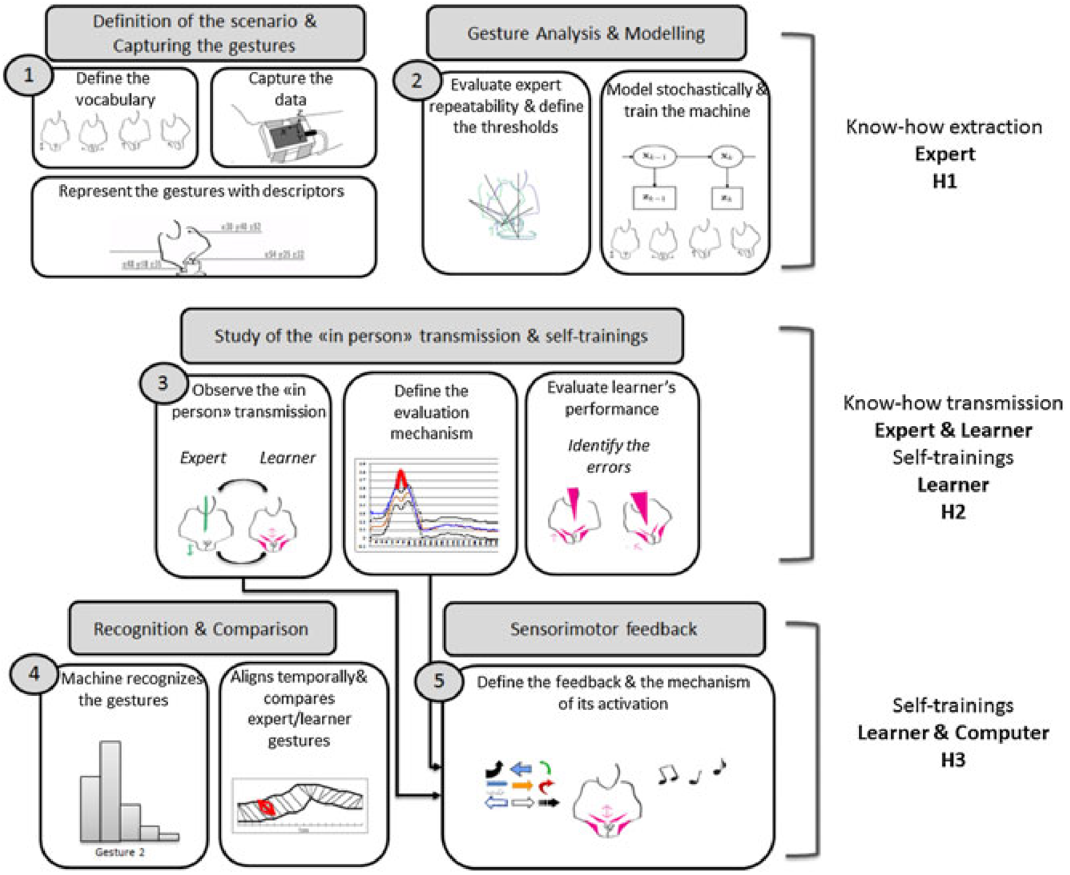

Definition of the scenario & Capturing the gestures
1
Define the vocabulary
Capture the data
Represent the gestures with descriptors
Gesture Analysis & Modelling
2
Evaluate expert repeatability & define the thresholds
Model stochastically & train the machine
Know-how extraction
Expert
H1
Study of the «in person» transmission & self-trainings
3
Observe the «in person» transmission
Expert
Learner
Define the evaluation mechanism
Evaluate learner's performance
Identify the errors
Know-how transmission
Expert & Learner
Self-trainings
Learner
H2
Recognition & Comparison
4
Machine recognizes the gestures
Gesture 2
Aligns temporally & compares expert/learner gestures
Sensorimotor feedback
5
Define the feedback & the mechanism of its activation
Self-trainings
Learner & Computer
H3

FIGURE 1 The global methodology for enhanced learning of gestural skills through gesture recognition technology [Colour figure can be viewed at wileyonlinelibrary.com]

(Piaget, 1976). Through interviews and observations (instrument used for this step), the natural interpersonal interaction is transformed into a schema where the learner plays passive or active roles, uses his or her senses and his or her perception. The evaluation mechanism, based on the threshold defined before, is developed at this stage and used for the evaluation of the learner's performance. Learning difficulties and errors are thus identified.

The fourth step in our methodology compares the expert's gesture model with the learner's gesture performed in real time, using machine learning techniques (HMM and DTW). It temporally aligns the two gestures and permits the identification of the main differences. During the alignment, the evaluation mechanism is activated: it checks whether learner's gesture's parameters are within the threshold defined. This comparison provides input for the mechanism of sensorimotor feedback activation for the assistance of the learner (*learner–computer*). In this methodological step, HMMs are used to recognize the gesture performed by the learner, because the gesture vocabulary may include more than 1 gesture. And the DTW is used for the temporal alignment of two time series (Manitsaris et al., 2014).

Finally, the aim of the last step was to define the sensorimotor feedback that consists of optical and sonic indications, implicit, or explicit that help the learner to adjust his deviations and errors. They are designed on the basis of the conclusions from the study of the in-person transmission conducted through observation and interviews. These indications are activated by the evaluation mechanism. In this paper, we describe the results from the application of this methodology in wheel-throwing pottery.

5 | APPLICATION OF THE METHODOLOGY IN WHEEL-THROWING POTTERY

5.1 | Extraction of pottery gestural know-how

In order to study if expert gestures in wheel-throwing pottery can be captured, analysed, and modelled we conducted an experiment with the initial participation of two potters, as described in a previous paper (Manitsaris et al., 2014). The scenario selected was the creation of a simple bowl (18/23 cm diameter) and the gesture vocabulary defined contained four or six gestures (G1 for the first gesture, etc.), depending on the object size and the quantity of clay used. Each gesture has been repeated and captured with inertial sensors 5 times. Consequently, the raw data were normalized, and the appropriate descriptors were used.

One problem encountered is that inertial sensors provide information about segment rotations expressed in angles that are difficult to interpret in the learning process. For example, informing the learner that his hand was lower than that of the expert's at a precise moment and differed by 32° cannot be easily assimilated from a pedagogical point of view. For this reason, we completed the technological set-up used for the experiment by implementing the depth camera Kinect to provide joint positions in 3D space for nine articulations of the upper body. The same four gestures, performed by a third potter, were thus recorded with the camera. The difference, in comparison with the previous recordings, consisted of the fact that here, the gestures had been performed virtually (Figure 2), with the wheel placed to the potter's left. This is a different wheel-throwing technique that facilitates data acquisition and permits the avoidance of occlusions; thus, the four gestures performed are the same for all three potters.

5.2 | Learning wheel-throwing pottery

5.2.1 | From in-person transmission to self-training without assistance

After the identification of expert gesture models, we proceeded to the observation of the in-person transmission. The learning process starts with the active role of the expert, who performs the gestures in real conditions with clay, and the more passive role of the learner who receives and interprets the visual information by observing the master potter (the sense used is *vision*). In parallel, the learner tries to imitate the presented four gestures virtually with the same rhythm and speed as the master potter. Then the roles are reversed and the expert observes the learner performing the gestures. She or she assists the learner by providing him/her with oral and sometimes visual observations (the senses used are *hearing and vision*). Finally, the learner starts performing the gesture in real conditions (Figure 3). The potter here still uses oral instructions such as "… push the clay higher, press the clay to centre it, close your hands to get a smaller object diameter" and also physical contact, to guide the learner in case of errors (the senses used are *hearing and touch*). During the in-person transmission, the learner is assisted and evaluated in real time by the potter.

After the in-person transmission, the learner is invited to train by himself or herself. In pottery, as well as in other manual arts and professions, the gestural skills can be acquired only through practice and experience. In this self-training step, 11 learners at beginner levels participated in our experiment (average age 26.2 years old; 10 right-handed and 1 left-handed; 6 women and 5 men; high degree of familiarity with technological devices such as PCs and smartphones but no previous experience in using interactive learning systems). In order to evaluate their performance, we used the same technological set-up that had been used for the capturing of Potter 3: Inertial sensors provided segment rotations, and the depth camera provided joint positions.

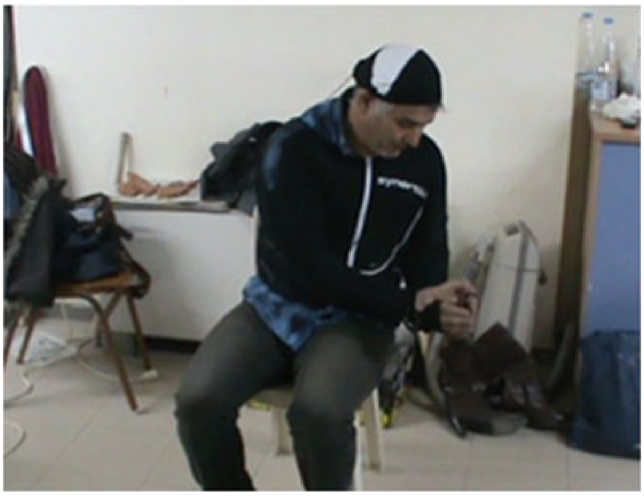


FIGURE 2 Potter 3 performs the gestures virtually [Colour figure can be viewed at wileyonlinelibrary.com]

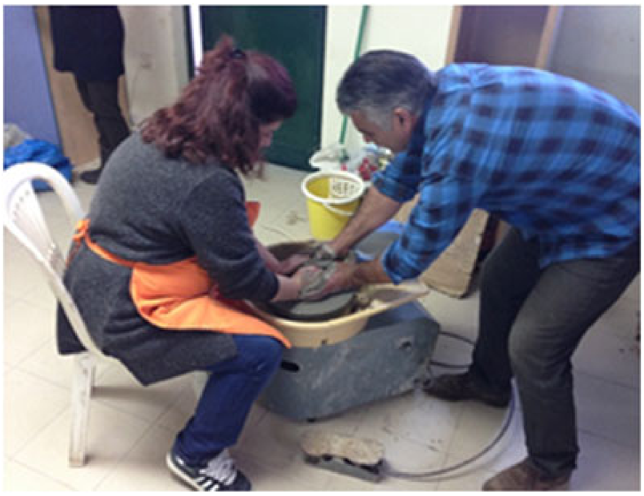


FIGURE 3 Potter 3 adjusts the learner's gestural position through physical contact [Colour figure can be viewed at wileyonlinelibrary.com]

5.2.2 | Learning wheel-throwing pottery with the assistance of sensorimotor feedback

On the basis of the observations from the study of in-person transmission, expert instructions can be categorized depending on (a) their function that is often linked to the moment they intervene and (b) the learner senses they activate. Explicit instructions give a clear guiding message ("… open your hands"), whereas implicit instructions require interpretation by the learner.

The design of the feedback provided by the computer is based on the expert instruction typology presented in Table 1 and on the principles of expert/learner relations which result from the observation of in-person transmission. We try to create metaphors between the natural potter–apprentice interaction and the learner/computer–human/machine interaction. In our methodology, we focus on correction and evaluation feedback because these have the greatest impact on the learner's execution of gestures and on the learner's motivation. The feedback provided focuses on the learner's hand distance deviation from that of the expert, as measured by the evaluation mechanism. We use both implicit and explicit optical feedback to support the learning process. More precisely, the learner's deviations are calculated and visualized in real time, whereas the learner is performing the gesture. The implicit visualization consists of curves/waves that vary depending on the deviation value so that the greater the deviation, the longer the

TABLE 1 Expert instruction typology

Prevention	Correction		Evaluation
Acoustic	Optical	Acoustic	Acoustic
Explicit	Implicit	Explicit	
		Implicit	

curve. The goal of the learner is to have a thin deviation line, equating with zero. For example, when the wave concerning horizontal distance deviation appears on the right, it means that the learner's hand distance is greater than that tolerated and she or he must "close" their hands to reduce the distance. This feedback is considered to be implicit because it requires the learner's interpretation of the curves. It does not indicate the decision that has to be made in order to correct the error. At the same time, more explicit optical instructions were implemented, showing the general distance trajectories to be performed, as we can see below in Figure 4.

The learner is also informed about the deviation through the use of acoustic feedback. The sound of a bell is used to attract the learner's attention and to make the learning process more entertaining. Finally, at the end of the gesture, evaluation feedback in also given to the learner, to provide him/her with a global picture of his/her performance, in order to reinforce his/her motivation. This takes the form of a global score that is given to the learner based on his or her temporal success percentage (100% minus percentage of temporal failure).

6 | EVALUATION

6.1 | Results of gesture modelling

To validate the given hypotheses, the Jackknife cross validation method was used. One repetition of each gesture trained the system, and the others were used for recognition. This process is applied in a circular way, until all the gestures to be used for training and recognition have been gathered. The basic idea behind the jackknife variance estimator lies in systematically recomputing the statistic estimate, leaving out one or more observations at a time from the sample set. To evaluate the machine's ability to recognize gestures, precision and recall metrics were used (Abdi & Williams, 2010). Precision metrics constitute the recognition rate that takes into consideration any erroneous correspondence to a model, whereas the system of recall metrics takes into consideration any missed recognitions.

$$\text{Precision} = \frac{\#\ \text{of}\ \text{True\_Recognitions}}{\#\text{of}\ \text{True\_Recognitions} + \#\ \text{of}\ \text{False\_Recognitions}}$$

$$\text{Recall} = \frac{\#\ \text{of}\ \text{True\_Recognitions}}{\#\text{of}\ \text{True\_Recognitions} + \#\ \text{of}\ \text{Missed\_Recognitions}}$$

TABLE 2 Recognition accuracy

	Precision (%)	Recall (%)
Potter 1 (rotations of 11 segments)	100	100
Potter 2 (rotations of 11 segments)	96	97.5
Potter 3 (positions of 9 articulations)	93.5	93.7

As we can see from the high recognition accuracy, shown in Table 2, the system is able to recognize different expert gestures and confirms that the kinematic aspects of potters' gestures can be captured and modeled. It also means that the models used for training are sufficiently from each other and that the experts have a high level of repeatability.

This repeatability had to be further evaluated in order to define the tolerance threshold that was necessary from the pedagogical perspective; however, this threshold could not be concerned with all the joints and articulations that could be captured with our equipment. Thus, in order to identify the body parts that play the most important role in the execution of pottery gestures, following the example of Volioti, Manitsaris, and Manitsaris (2014), we applied a principal component analysis to the dataset recorded with the camera in order to reduce the data dimensions. In addition to the positions of nine articulations on three axes, we used the variability of hand distance, because the expert mentioned that it is an important parameter in learner evaluation. In Figure 5 below, the variables with the highest values on the vertical axis are related to the effective pottery gestures that directly influence object creation, whereas variables with values on the horizontal axis belong to the accompanying secondary gestures. As we can see from this analysis, the expert's statement is confirmed because

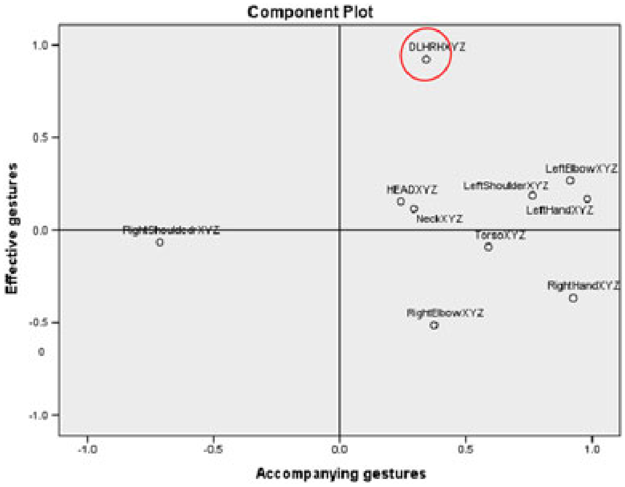


FIGURE 5 The component plot of the principal component analysis applied to pottery gestural data [Colour figure can be viewed at wileyonlinelibrary.com]

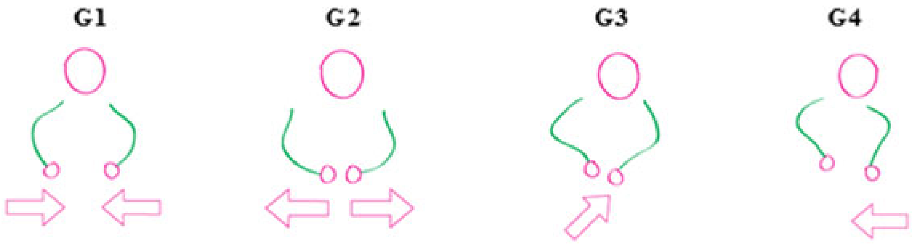


FIGURE 4 The four indications used as explicit optical feedback [Colour figure can be viewed at wileyonlinelibrary.com]

hand distance (DLHRHXYZ) appears to be the variable that constitutes the principal component in the dataset.

After the identification of the most important gesture variables, we could proceed to the evaluation of repeatability through the calculation of the distance standard deviation (*SD*) of the third potter. For this, first of all, we needed to temporally align the five repetitions of each gesture with the use of DTW technique (Manitsaris et al., 2014). Then we calculated the *SD* for each time frame and each axis, and we took the maximum value as the tolerance threshold (λ). After the transformation of these values into centimetre, we can see the threshold in the following table.

We can conclude from this Table 3 that, in general, Potter 3 has a very high level of repeatability. In other words, this table shows that while executing G2 twice, the potter may permit himself to have a difference of 1.02 cm on axis *X*. The *Z* axis has the most important deviation from one repetition to the other, and this may be caused by the potter moving his chair or changing his position in front of the camera. Consequently, axis *Z* will not be used for the evaluation of the learning process.

Thus, the application of the described methodological steps and the evaluation results provide us with confirmation of the scientific hypothesis that GRT can contribute to the capturing and modelling of kinematic aspects of expert technical gestures. The high recognition accuracy, achieved here, demonstrated the success of gesture modelling and the statistical analysis (principal component analysis, *SD*) permitted us to identify the important characteristics of this precise gestural know-how.

6.2 | Mechanism for learner evaluation

In order to understand the learner's performances, the two parameters that were evaluated were those of the learner's (a) temporal deviation and (b) spatial deviation. The first concerns the duration of the learner's gestures that should be as close as possible to that of the expert. According to the expert, the learner should assimilate the speed and rhythm of each gesture and the acceptable temporal deviation is that of 5 s. In order to help the learner place his performance within a temporal framework, we implemented a second counter. In order to calculate learner's spatial deviation, an evaluation mechanism was proposed, which was based on machine learning. Because we needed to compare the two gestures performed by different persons, we needed to align their duration. Expert gestures are thus used one by one for the training of gestural models, and the learner's data is used for recognition. For the calculation of learner's deviations, we use the tolerance threshold (λ) that is added to expert Euclidian hand distance. Every time the learner's hand distance (DL) on axis *X* or *Y* exceeded the greatest tolerated expert distance (DE + λ) or was less than the least tolerated distance (DE-λ), his or her performance was considered to deviate from the model.

$$\begin{aligned}
&\text{if } DL_x<(DE_x-\lambda) \text{ then } (DL_x-(DE_x-\lambda))\\
&\text{if } DL_y<(DE_y-\lambda) \text{ then } (DL_y-(DE_y-\lambda))\\
&\text{if } DL_x<(DE_x+\lambda) \text{ then } (DL_x-(DE_x+\lambda))\\
&\text{if } DL_y<(DE_y+\lambda) \text{ then } (DL_y-(DE_y+\lambda))
\end{aligned}$$

To calculate the total deviation of the gesture, instant deviations were added, giving a final number of one per axis. In order to have a clearer image of the deviation, in the analysis below, we calculated the average value of these two deviations.

Moreover, for a comparative measure between expert and learner gestures, we used recognition accuracy, which depended on their degree of similarity. When the input data used for recognition presented high probabilities for correspondence with the reference gestures, then the likelihood measure was high.

6.3 | Temporal deviation with and without computer assistance

During the self-training sessions, the learner tried to perform the gestures according to the motor memories she or he acquired with the in-person transmission. The gesture duration was approximate. While using our system, the learner received the information about expert duration and his current duration and could adapt it. To evaluate temporal progress, we compared the sum of the difference between the average expert (DurEx) and learner duration (DurL).

$$\sum_{i=0}^{i=11}\left(\overline{\mathrm{DurL}_i}-\overline{\mathrm{DurEx}_i}\right)$$

We can conclude from Table 4 that the sum of duration average difference between the expert and the learner was reduced for two of the four gestures. In the case of G2 and G3, the indication of current and desirable duration permitted the learners to approach the temporal goal. IT also shows that the gesture with the biggest deviation is G4. This could be explained by the fact that these gestures contain an important number of subgestures of great amplitude (taking a wire, removing the object from the wheel etc.). The fast coordination of these substeps requires experience and cannot be achieved at beginner's level.

In addition, according to the expert, (a) the difference in average duration between the learner and the expert per gesture should be lower than 5 s and (b) the acceptable temporal deviation between the different repetitions of the same gesture by the learner should be 5 s or less. Table 5 presents the number of learners per gesture with the average learner duration surpassing the expert duration by more than 5 s, with and without feedback.

TABLE 3 Tolerance threshold (λ) in centimetre for each gesture

	X	Y	Z
G1	1.7 cm	1.9 cm	3
G2	1.02 cm	1 cm	3.42
G3	1.47 cm	1.13 cm	2.8
G4	1.71 cm	1.6 cm	2.9

TABLE 4 The sum of duration average difference in second between the expert and the learner for 4 gestures with and without feedback assistance

	G1	G2	G3	G4
Without feedback	54.5	31.94	36.01	90.78
With feedback	56	24.9	27.9	90.8

TABLE 5 Number of learners per gestures with $\left(\overline{\mathrm{DurL}_i}-\overline{\mathrm{DurEx}_i}\right)>5$ s

	G1	G2	G3	G4
Without feedback	6	2	3	8
With feedback	4	1	2	10

Table 5 shows that the number of learners who have important deviations from the expert's average has been reduced for three of the four estures. In the case of G4, two more learners′ averages deviate from the acceptable temporal threshold, while using the feedback. This is also linked to the particularities of this last gesture. It seems that during the learning process, the articulation of the different substeps of a gesture (G4) requires additional time.

The expert also states that it is important for the learner to have temporal homogeneity while repeating the same gesture; however, we can observe that several times the learner may perform the same gesture with a very different duration (from 25 to 45 s or even 50 s for the G2, etc.). Table 6 shows the number of learners per gesture who have a standard deviation between the different repetitions of the same gesture of greater than 5 s.

As we can see from Table 6, with the use of sensorimotor feedback, *SD* was reduced for three of the four gestures. In particular, for G1 and G2, all the learners had a temporal deviation per gesture <5 s, which is the desirable result according to the expert. We can thus conclude that sensorimotor feedback, and, more precisely, the time counter, helped learners reach a better perception of the desirable duration.

6.4 | Spatial deviation with and without computer assistance

6.4.1 | Global performance and general deviation of four gestures

If we evaluate the hand distances of the 11 pottery learners who participated in our experiment and compare them with the results of self-training with and without computer assistance, we observe that, in sum, the learners′ kinematic performance improved with the use of sensorimotor feedback. According to the sum of learners′ deviations, three of the four total gestures were improved. The interpretation of visual and sonic feedback seems to have helped students to understand the correct gesture trajectories. In contrast, as indicated in Table 7, the performance of G4 improved for a minority of learners, which is due to the different nature of this gesture, as explained previously.

TABLE 6 Number of learners per gesture with *SD* > 5 s

	G1	G2	G3	G4
Without feedback	3	4	3	2
With feedback	–	–	1	2

6.5 | Gradual progress in the learner′s performance through repetition and guidance from feedback

After the general confirmation of the contribution of feedback to the learning process, we focused on its role in the learner's gradual progress, from the first to the last repetition. For this, we calculated the number of learners who had their best performances at the beginning (first three repetitions) or at the end (last two) of the training process. The best performances were those with the least spacial deviation of hand distances in comparison to the expert model.

As can be seen from Table 8, an important number of learners had their best performance at the beginning of the training, with a regressive tendency for the last repetitions. This could be due to different degrees of attention, concentration, and fatigue between the first and the last repetition. In contrast, the number of learners with their best performances at the end of the training process increased for three of the four gestures, with the use of sensorimotor feedback. This seems to confirm the fact that (a) learners need time to assimilate the sensorimotor feedback function and (b) it contributes to their concentration and motivation, because they try to improve their performance by the end of the training process.

To give more precise examples, we offer the performance evolution of two learners for G3. As can be seen in Figure 6, L1 has regressed when performing without computer assistance, but this tendency is reversed with the feedback. Here, only one repetition (the 3rd) presents an important deviation. The same phenomenon is observed for L3, who managed to improve his deviation, with one repetition (the 2nd), following 15 cm of deviation. These numbers illustrate the fact that the learners needed to experiment with the gestures and with computer assistance. They intentionally provoked

TABLE 7 The average deviation on the *X* and *Y* axes in centimetre, for each learner, for four gestures, with (F) and without feedback (NO F)

	G1 NO F	G1 F	G2 NO F	G2 F	G3 NO F	G3 F	G4 NO F	G4 F
L1	97.15	79.59	51.49	64.82	29.71	26.78	107.92	92.11
L2	208.68	152.35	103.72	05.03	112.77	39.17	96.97	166.90
L3	185.88	49.71	09.54	02.41	17.34	23.20	192.34	100.01
L4	96.25	29.73	19.36	11.81	38.22	12.61	133.74	338.45
L5	59.55	55.45	28.86	08.56	133.38	34.13	53.09	227.67
L6	233.13	72.60	192.11	03.10	235.73	32.18	328.70	93.30
L7	82.82	35.43	78.69	03.22	24.55	03.13	181.98	491.83
L8	312.75	32.84	74.83	10.18	213.42	45.17	87.96	184.67
L9	122.80	94.89	69.93	25.75	30.41	21.04	92.44	135.90
L10	170.09	46.70	05.30	07.96	25.80	47.56	328.56	223.89
L11	134.53	98.72	05.99	13.14	65.65	30.03	258.89	103.54
Σ	**1703.62**	**748.01**	**639.82**	**155.97**	**926.98**	**315.00**	**1862.58**	**2158.25**

TABLE 8 Number of learners with the best performances at the beginning or end of the training process, with (F) and without (NO F) feedback

	G1	G2	G3	G4
First 3 repetitions NO F	4	5	9	8
Last 2 repetitions NO F	7	6	2	3
	G1	G2	G3	G4
First 3 repetitions F	5	2	2	6
Last 2 repetitions F	6	9	9	5

TABLE 9 Precision (P), recall (R), and the total (TP, TR) of gestures performed with feedback (left), and without (right)

	M1	M2	M3	M4	R (%)	TR (%)
G1	53	2	—	–	91	**80**
G2	8	42	4	1	67	
G3	—	17	35	3	65	
G4	1	—	—	54	98	
P	81%	70%	77%	93%		
TP	**80%**					
	M1	M2	M3	M4	R (%)	TR (%)
G1	53	1	—	1	96	**89**
G2	—	55	—	—	100	
G3	—	18	37	—	67	
G4	3	1	—	51	93	
P	95%	72%	100%	98%		
TP	**91%**					

kinematic errors in order to observe the system′s reaction, and they needed a certain amount of time for adaptation in order to learn how to interact with the machine.

6.6 | Evaluation of gesture recognition accuracy

At this stage, we can speculate that any improvement in the recognition accuracy of the learners′ gestures captured during the computer-assisted (feedback) self-training sessions would seem to demonstrate an improvement in performance. The machine′s greater ability to recognize observations (learners′ gestures in G1, etc.) would seem to indicate that they are closer to the models (expert gestures —M1, etc.) used in training.

Table 9 below presents the results of learner gesture recognition performed without feedback. Because each gesture was repeated and captured 5 times with 11 learners, we have 55 instances of gesture performance in our dataset. In Table 9, we can see the correspondence of gesture error recognized and the model that has been assigned by the system (e.g., G1–M2). The gesture with the lowest recall is the third (17 repetitions have been assigned to M2). This may be due to the fact that both G2 and G3 are performed within the same spatial framework (wheel-throwing diameter) without many subgestures. The total precision and recall reached was 80%.

Interestingly, when we trained the hybrid machine learning system with the same expert gestures, creating the same four models, but learners′ gestures performed with feedback assistance were used for recognition, we could observe better recognition accuracy. More precisely, G2 recall and M3 precision improved respectively from 67% and 77% to 100%. This improvement could be linked with the better spatial performance of gestures, as explained in the previous section. Most importantly, total precision and recall increased by approximately 10%, reaching 90%.

The analysis of learners′ gestures, performed without feedback, permitted the identification of temporal and spatial deviation in comparison to expert gestures. Motion capture, GRT and machine learning techniques allowing the temporal alignment of two datasets, permitted quantitative evaluation of learners′ gestures in wheel-throwing pottery, thus providing confirmation of the second hypothesis in this

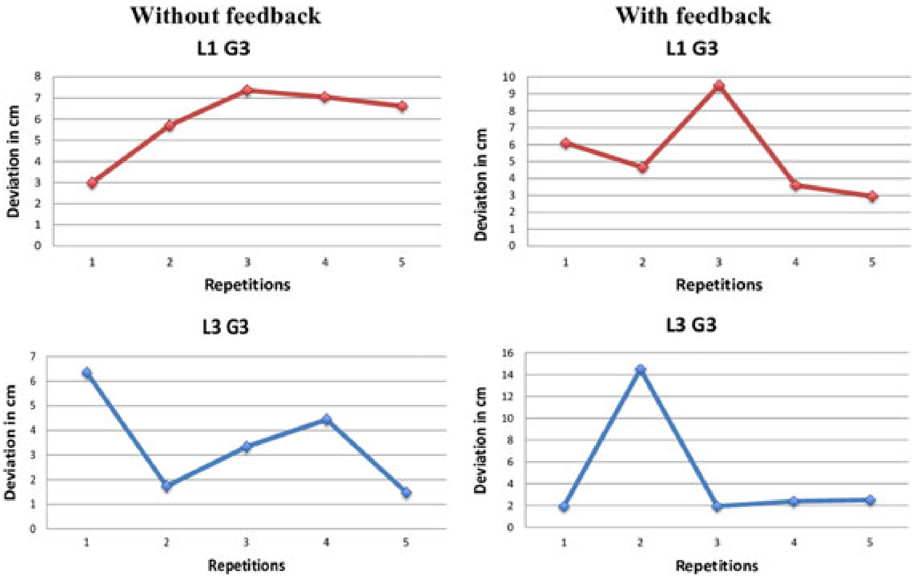


FIGURE 6 Examples of average sum deviation on the *X* and *Y* axes for five repetitions by Learners 1 and 3, in centimetre [Colour figure can be viewed at wileyonlinelibrary.com]

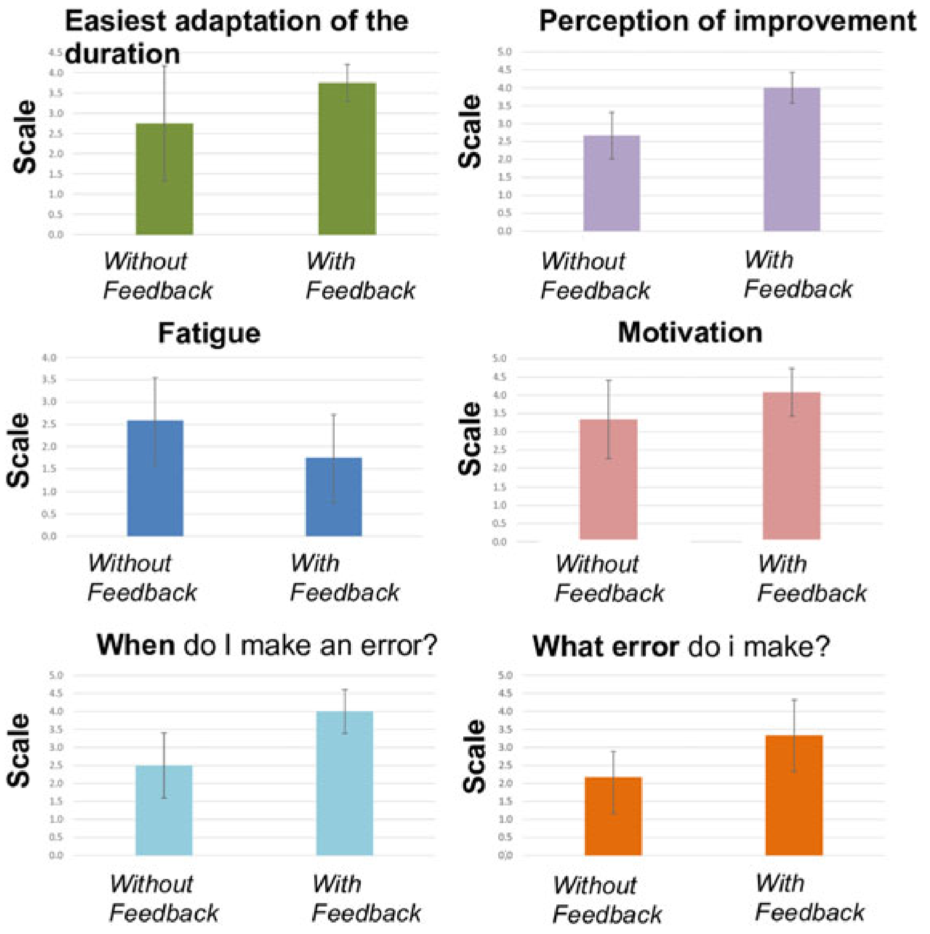


FIGURE 7 Qualitative evaluation results [Colour figure can be viewed at wileyonlinelibrary.com]

research. The reduction in temporal and spatial deviation of gestures performed with feedback assistance and the improvement of their accuracy recognition, validates the third and final hypothesis that sensorimotor feedback assists with self-training and contributes to improving learners′ performance.

6.7 | Qualitative evaluation of gestures learning and system′s usability

In parallel with the quantitative evaluation of gestures learning, we implemented a qualitative one, through questionnaires containing open and semiconductive questions. The goal of the questions was to collect learners′ global impressions about the system, the feedback, and the whole learning process with and without the use of computer assistance. To evaluate the different parameters such as fatigue and interaction, we used a rating scale from 0 to 5, from *a weak* to *a strong degree* of perception. The graphics of Figure 7 present the average rate per question together with the standard deviation of 11 learners who participated in the experiment. The results presented in the figure correspond to questions: “How easy it was to adapt your gesture′s duration to the one of the expert (with and without feedback)?;” “From the 1st to the last repetition, how easy it was to perceive an improvement of your performance (with and without feedback)?;” “How strong was the fatigue /motivation you felt while repeating the gestures (with and without feedback)?;” “How easy it was to perceive the moment you were making an error and what error it was (with and without feedback)?” According to them, when using our system with the feedback, they feel less tired after the self-trainings than without it. Their motivation increased on average by 1 unit and passed from moderate without feedback (3.2) to strong with feedback (4). Also students seem to perceive positively the duration counter and confirm that it helps to put their gestures in the desired time frame. Learner′s perception of their gestures improvement through repetitions seems to be more important with the use of feedback (from 2.6 to 4). From one iteration to the other, students have the opportunity to compare both their score and the feedback (How many times the sound was activated?, How thick was the deviation line? etc.). This permits to better follow the progress of learning. According to the questionnaire, the feedback also reinforces a better perception of the moment the learners are making an error and of what error do they make (bigger or smaller hands distance).

7 | DISCUSSION AND CONCLUSION

In this paper, we presented a methodological framework for the modelling and transmission of gestural skills. The system proposed, based on machine learning techniques, appears to support self-training through the assistance it provides to the learner. It is able to calculate the learner′s kinematic errors and provide feedback, in real time, guiding them and permitting them to adjust the gestures and overcome errors. In contrast to the self-training without computer assistance, continuous evaluation makes the learning process more interactive and thus reinforces the learner's motivation. The design of the feedback and its activation mechanism was inspired by natural expert–learner relations and interaction.

This study still presents some limitations because we focus mostly on the kinematic aspects of wheel-throwing pottery gestures, whereas kinetics (pressure, force measurement, etc.) and finger gestures are also important parameters to be evaluated. Collecting and analysing finger gestures data constitute a real challenge because it is not possible to use intrusive technologies and influence the creation process. At the same time, the use of nonintrusive computer vision techniques is also problematic because of occlusion issues; it is difficult to extract the fingers from the scene, when they touch the clay. The integration of kinetic gesture aspects and fingers motion into our methodology is a priority for our current and future research topics. The other limitation is linked to the fact that our methodology has been evaluated with virtual gestures. It is important to apply and evaluate it with gestures performed in real conditions with the use of clay. However, as mentioned before the existing noninvasive motion capture technology cannot presently provide sufficient reliable data.

As we have seen in literature review, multimedia technologies are traditionally used for the preservation and transmission of gestural know-how (Chevallier, 1991; Bril, 2011; Wang et al., 2011; Kim, 2011). Some projects make use of motion capture technologies to collect a more complete gestural performance dataset (Rasamimanana & Bevilacqua, 2009), (Raptis et al., 2011). As we have mentioned in the literature review, multimedia technologies don't provide any precise information about the gestural performance, the recordings remain two-dimensional, whereas pedagogical contents based on multimedia don't permit the learner to actively interact with it. From the other hand, studies conducted with motion capture technologies required expensive or invasive equipment, and the feedback provided to the learner is based on simple tracking of body joints.

The goal of our approach is to provide a system that would be able to guide the learner in real time, whereas he is performing the gesture. The added value of our framework is the fact that it is based on low cost motion capture technology (Kinect camera) and that precise kinematic features are modelled with the use of machine learning techniques that permit also to compare in real time the expert gestural model and learner's performance. Biomechanical aspects of gestures are captured, modelled, and recognized, compared with the reference gesture to make the system able to provide a feedback, a guidance to the learner.

ACKNOWLEDGEMENTS

This work is funded by the European Commission (European Social Fund) via the i-Treasures project (Intangible Treasures—Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures FP7-ICT-2011-9-600676-iTreasures). It is an Integrated Project (IP) of the European Union's 7th Framework Program "ICT for Access to Cultural Resources."

ORCID

Alina Glushkova http://orcid.org/0000-0002-6214-2034

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., & Rasamimanana, N. (2009). Continuous realtime gesture following and recognition. In *International gesture workshop* (pp. 73–84). Berlin, Heidelberg: Springer.

Boyer, E. (2015). Continuous auditory feedback for sensorimotor learning. Unpublished doctoral dissertation, Université Pierre et Marie Curie-Paris VI, France.

Bril, B. (2011). Description du geste technique: Quelles méthodes? *Techniques & Culture*, *1*, 243–244.

Chevallier, D. (1991). Savoir faire et pouvoir transmettre. Transmission et apprentissage des savoir-faire et des techniques. Les Editions de la MSH.

Dale, E. (1969). Audiovisual methods in teaching.

Glushkova, A., & Manitsaris, S. (2015). Gesture recognition technologies for gestural know-how management: Preservation and transmission of expert gestures in wheel-throwing pottery, Proceedings of CSEDU Conference, Lisbon, Portugal, 23–25 May.

Godbout, A., & Boyd, J. E. (2010). Corrective sonic feedback for speed skating: A case study. Proceedings of the 16th International Conference on Auditory Display, pp. 23–30

Herzig, N., Moreau, R., & Redarce, T. (2014). A new design for the BirthSIM simulator to improve realism. Proceedings of the Engineering in Medicineand BiologySociety (EMBC), 36th Annual International Conference of the IEEE (pp. 2065–2068).

Jégo, F., Paljic, A., & Fuchs, P. (2013). User-defined gestural interaction: A study on gesture memorization. IEEE 3D User Interfaces 3DUI 2013, Orlando, Fl., United States.

Kim, E. S. (2011). Choreosave: A digital dance preservation system prototype. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–10.

Le Bellu, S., & Le Blanc, B. (2010). How to characterize professional gestures to operate tacit know-how transfer? *Electronic Journal of Knowledge Management*, *10*(2), 142–153.

Malamed, C. Get smart about designing learning experiences: The power of interactive learning, http://theelearningcoach.com

Manitsaris, S., Glushkova, A., Bevilacqua, F., & Moutarde, F. (2014). Capture, modeling and recognition of expert technical gestures in wheel-throwing art of pottery. *Journal on Computing and Cultural Heritage (JOCCH)*, *7*(2), 10.

Newell, K. M. (1991). Motor skill acquisition. *Annual Review of Psychology*, *42*(1), 213–237.

Ng, K. C., Weyde, T., Larkin, O., Neubarth, K., Koerselman, T., & Ong, B. (2007). 3d augmented mirror: A multimodal interface for string instrument learning and teaching with gesture support. Proceedings of the 9th International Conference on Multimodal Interfaces (pp. 339–345), ACM.

Piaget, J. (1976). *Piaget's theory*. Berlin & Heidelberg: Springer.

Raptis, M., Kirovski, D., & Hoppe, H. (2011, August). Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation* (pp. 147–156). Vancouver: ACM.

Rasamimanana, N., & Bevilacqua, F. (2009). Effort-based analysis of bowing movements: Evidence of anticipation effects. *Journal of New Music Research*, *37*(4), 339–351.

Rieber, L. P. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research and Development*, *44*(2), 43–58.

Schank, R. (1997). Virtual learning. McGraw-Hill Companies.

Volioti, C., Manitsaris, S., & Manitsaris, A. (2014). Offline statistical analysis of gestural skills in pottery interaction. Proceedings of the 2014 International Workshop on Movement and Computing, ACM.

Wang, K. A., Liao, Y. C., Chu, W. W., Chiang, J. Y. W., Chen, Y. F., & Chan, P. C. (2011). Digitization and value-add application of bamboo weaving artifacts. In *Digital libraries: For cultural heritage, knowledge dissemination, and future creation* (pp. 16–25). Berlin Heidelberg: Springer.

Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience*, *12*(12), 739–751.

How to cite this article: Glushkova A, Manitsaris S. Gesture recognition and sensorimotor learning-by-doing of motor skills in manual professions: A case study in the wheel-throwing art of pottery. *J Comput Assist Learn*. 2018;34:20–31. https://doi.org/10.1111/jcal.12210

A tabletop instrument for manipulation of sound morphologies with hands, fingertips and upper-body.

Edgar Hemery
Center for Robotics - MINES
ParisTech, PSL Research
University
60, Bd Saint-Michel
75272 Paris, France
edgar.hemery@mines-paristech.fr

Sotiris Manitsaris
Center for Robotics - MINES
ParisTech, PSL Research
University
60, Bd Saint-Michel
75272 Paris, France
sotiris.manitsaris@mines-paristech.fr

Fabien Moutarde
Center for Robotics - MINES
ParisTech, PSL Research
University
60, Bd Saint-Michel
75272 Paris, France
fabien.moutarde@mines-paristech.fr

ABSTRACT

We present a musical instrument, named the *Embodied Musical Instrument* (EMI) which allows musicians to perform free gestures with the upper–body including hands and fingers thanks to 3D vision sensors, arranged around the tabletop. 3D interactive spaces delimit the boundaries in which the player performs metaphorical gestures in order to play with sound synthesis engines. A physical-based sound synthesis engine and a sampler have been integrated in the system in order to manipulate sound morphologies in the context of electro-acoustic and electronic composition.

Author Keywords

Gesture data recording; Gesture Recognition; Physical-based sound synthesis; Morphological transformations; Sound image

ACM Classification Keywords

H.5.m Information Interfaces and Presentation Interaction styles, H.5.5 Information Interfaces and Presentation Sound and Music Computing, J.5 Arts and Humanities Performing arts .: Miscellaneous

INTRODUCTION

Theorising the concept of sound objects in the *Traité des objets Musicaux* [1], Pierre Shaeffer initiated a form of electro-acoustic music, called *musique concrète*, which suggests the listener to identify the sounds individually and to value their sound morphology equally as rules of melody, harmony, rhythm, metre, etc. By the means of new recording and broadcasting technologies, Pierre Shaeffer made tape collages of sound recordings and started exploring music through textures of sounds. This approach encouraged composers to make use of any sound materials they could get a hand on. This idea followed closely the Futurists' manifest, the Art of Noise [2] which first put forward the use of machines sounds for compositional purposes and Edgar Varèse who in 1914, was using "the musical matter itself" in his compositions. Later, with the birth of computer analysis of sound spectrograms, theories of concret music evolved in new musical trends such as spectral music and other musics which focus on timbre as an important element of structure or language. It is worth recalling that before these new perspectives raised in various forms, western music was focused on pitch structures (harmony, modality), construction of musical forms (themes, motives), and rhythm (meter). Timbre was simply used as a matter of colorisation of musical structures and considered in terms of orchestration. Furthermore, electro-acoustic music composition can even be regarded as painting or sculpture [3] where the artist works with shapes and textures.

In 1913, the italian Futurist Luigi Russolo, wrote in *The art of noise*: *'It will be through a fantastic association of the different timbres and rhythms that the new orchestra will obtain the most complex and novel emotions of sound'*. In 1916, Russolo reported police intervention to stop riots at his concert. In the 1950's, Varèse's piece *Deserts*, provoked bad reactions in the audience because of its absence of theme and melodic structures. This facts shows that either the people were not ready for new kinds of experimental sounds or that the music was irritating. Truth is that in the 1930's, composers had only the crudest control over the sounds they were using. Noise music was only at its beginning and artists did not have appropriate tools for controlling it, creating a distance between the composer and the musical manipulation. Varèse and Cage work on percussion music was a natural step in the long process of admitting unpitched sounds into music. It required several generations until people could identify themselves to certain kind of sounds. Nostalgia and melancholia for instance were difficult to convey with these early electronic music, dissociating music to humans' emotions.

In the mean time, the first electronic music instrument appeared in 1928 with the world famous eponymous creation of the russian inventor Leon Theremin. Interestingly, this invention, prior to the computer and motion sensors, was based on *air-gesture* capture as its working principle. However, this latter invention was mostly used to play a classical music repertoire and was not integrated in electro-acoustic compositions at that time. Elektronische music resulted in the 50's

DOI string from ACM form confirmation

in Cologne from the research of composers such as Stockhausen, Eimert, Beyer and Eppler's on sound synthesis. It was a radically different approach to the concret music since the music and sounds were entirely produced by electronic means. For Eimert, sound synthesis was a real musical control of nature based on the use of sine tones as the fundamental of the art. The main interest of electronic sound synthesis was a desire to control over every aspects of musical compositions. But if this technique changed entirely the course of music, it temporarily lead to total determinism and formalism in the compositional approach.

As mentioned before, one of the main problem in electro-acoustic and electronic music is the distance between composers and the composing medium, which later became the computer and the interface. As Pierre Shaeffer wrote: '*The lack of intentional control over musical affect, together with the fact that compositions emanating from such a wide range of compositional aesthetics all produced the same impressions, implicate the common rudimentary sound manipulation technologies*'. This inadequacy for sound manipulation prevents the spontaneity and the emotional intention of the musician. Later in the 80's, improvements over computers CPU capacities allowed for real time control of sound synthesis. This way, the composer had access to a very wide range of sounds and could trigger them spontaneously with the help of keyboards, cursors, mixing desks, buttons and track pads. As Emmerson [4] pointed out, ' *In the 1980s, two types of computer composition emerged from the studio to the performance space, one more interested in event processing, the other in signal processing*'.

Joel Ryan from the Steim Institute in Amsterdam wrote: *In order to narrow this relationship between technology and musicians, it is as much the problem in collaboration to get technologists to respect the thinking of the artists as it is to educate the artists in the methods of the technology* [5]. Signal processing as it is taught to engineers is guided by such goals as optimum linearity, low distortion, and noise. This goals may not be in accordance with musicians wishes. The temptation of programmers is to concentrate on the machine logic rather than the idea of the artist. New suggestions of sound techniques that fit musical expressivity are needed. Several research groups such as the IRCAM in Paris, the STEIM in Amsterdam, the CCRMA in Stanford, and the MTG of Pompei Fabra-Barcelona to name a few, have started to focus on system designs easing interactive manipulation of sounds. This includes thinking about intuitive interfaces, gestural controllers, communications protocols, network designs and an understanding of how they can all be interconnected.

STATE OF THE ART

In the past few years, non-intrusive movement and gesture analysis have been integrated in consumer electronics thanks to the progress made in 3D cameras technology and computer vision algorithms. Computer vision is a branch of computer science interested in acquiring, processing, analysing, and understanding data from images sequences. Non-intrusive gesture tracking systems are ideal for musical performances since they allow freedom in body expression, are not intrusive and are easy to calibrate. Therefore, a musical interface – or instrument – which draws gestural data from vision sensors, feels natural from the user's experience point of view, provided that gesture to sound mapping is intuitive and has a low latency response.

Performing arts have embraced this type of technology from its very beginning, seeing in it an extraordinary springboard for creation of new exciting interactions, highlighting the performer's body and gestures. Starting from 1982 with David Rokeby' series of performances with *Very Nervous System*[1], a new area of embodied interaction making use of cameras and computer vision algorithms was born. A global vision on this scene reveals that the Kinect and the Leap Motion, have already been popular choices among musicians and sound artists for live performances.

Recent advances based on the Leap Motion show its abilities to control high-level music control thanks to a 3D touch-like gesture. GECO is a Leap Motion app (available on the Leap Motion' market space *AirSpace*) which allows to control MIDI, OSC or CopperLan protocols with a simple 3D-gesture vocabulary. Han and Gold [6] use the Leap Motion to create an *air-key* piano and an *air-pad* machine drum while making use of the third dimension to control the sound intensity via the hand' velocity computation. The *BigBang rubette* [7] module uses the Leap Motion to control notes, oscillators, modulators or higher level transformations of sounds and/or musical structures. Alessandro et al. [8] and Silva et al. [9] combined respectively a Leap Motion and transparent sheet of PVC [8] and glass [9] in order to grasp finger movements occurring prior to the touch with the sheet. It is worth noticing that this last example somehow meets with multi-touch tablets and tabletops since the paradigm is based on a finger-screen contact.

Nowadays, multi-touch screens and Omni-Touch wearable interfaces [10] offer tangible interactions that are restricted to a flat surface with finger tapping, scroll, flick, pinch-to-zoom etc. (refer to [11] for extended reference guide of touch gestures). The ReacTable [12] has launched a multi-player tangible interaction of a unprecedented kind with real objects communicating on a multi-touch tabletop. The ReacTable' objects display images that are recognized by an infra–red camera, sending information about the type of sound to be generated to the system. It is striking how this approach falls in with the concept of sketches and shapes of sonic objects described in Schaeffer's typology [1]. A similar work by Thoresen [13] introduced a set of graphical symbols apt for transcribing electro–acoustic music in a concise score, simplifying the sometimes overwhelming complexity of Shaeffer' *Typo–Morphology*.

In the same vein, the use of extra objects such as digital pen in the music production app on Microsoft Surface tablet[2] gives very interesting and intuitive ways for achieving high-level sound control parameters such as drawing amplitude and filter envelopes. This smart tabletop, along with the ReacTable

[1] www.davidrokeby.com/vns.html

[2] http://surfaceproaudio.com/

discussed above, belong to the first generation of devices and instruments to allow embodiment and intuitive manipulation of sound objects.

In line with these latter examples, our instrument, that we are discussing in this article, is an interactive tabletop for playing music in 3D space where the upper–body and fingers' free–movements in mid-air extend the action of the fingers' physical contacts with the table. While Microsoft Research has developed similar technologies and set-ups for grabbing and manipulating 3D virtual objects on and above a tabletop surface with finger gestures ([14] and [15]), the EMI pushes ahead sound mapping strategies in the 3rd dimension. Additionally, the EMI is in line with extended piano-keyboard devices such as the Seaboard by Roli[3] and the TouchKeys[4] by Andrew McPherson, but brings the extended interaction to mid-air with both fingers and upper-body gestures thanks to 3D vision sensors such as the Leap Motions and the Kinect.

We start by describing the structural aspect of our instrument, the sensors that are used and the concepts of *micro* and *macro* bounding boxes articulated around the framework. The Section *Musical embodiment on a tabletop instrument* depicts the metaphors used while designing the interactions with the system. A variety of sound synthesis controls are presented in the section *Sound morphologies manipulation*, showing the musical capacity of our instrument to control sound morphologies. The section *Latency assessment of the EMI* presents a first latency assessment of the system. Then we conclude and give a view of our further works.

DESIGNING A FRAMEWORK

Structure of the instrument

The whole instrument is articulated around an acrylic sheet, which serves as a frame of reference for the fingers. The acrylic sheet is placed 10mm above two Leap Motions, where the sensors' field of view covers the area best and underneath a Kinect placed 1.20mm in front (see figure 1 and 2). The cameras are described later in this section. The sheet also constitutes a threshold of detection for the fingers: one triggers the sound by fingering the tables surface. Gestural interaction is not limited to this surface, but takes part inside a volume above the table. The tracking space serves as a *bounding box*, delimiting the sensors' field of view in which the data are robust and normalized.

The boundary engendered by the table's surface eases the repetition of a type of gesture. This conclusion raised from the difficulties of gesture repetition observed in *air*-instruments, where the movement is done in an environment with no tangible frame of reference. In this regard, it is a profitable constraint to add this surface since it enables the user to intuitively place his/her hands at the right place and helps repeating similar gestures.

As the *Embodied Musical Instrument* is to be used for both performances and learning contexts, it is portable, light, solid and foldable. We have conducted experiments, changing the tilt of the sheet to meet with the literature results concerning the wrist and shoulder posture during touch-screen tablet use [18]. However, movements with the EMI being wide and dynamic, wrist radial deviation was not constant enough to take into consideration optimal tilt angles for specific applications. At last, the sheet supports the arms and allow the user to rest, thus avoiding the *gorilla arm* effect which results in a fatigue while repeating gestures in the air [19].

Vision-based 3D sensors

We present here two types of vision-based sensors, which are used in our research. As this technological field is growing fast, we could not explore all the existing sensors possibilities; however, the sensors we have chosen are well documented, largely spread, low cost and fit our requirements. The first type of sensor is the Microsoft Kinect depth camera. The first version of the Kinect, along with the OpenNI skeleton tracking software delivers a fairly accurate tracking of the head, shoulders, elbows and the hands, but not fingers. It has a 43° vertical field of view, 57° lateral field of view and a ranging limit varying from 0.8 to 3.5 m. Its latency, around 100 ms and its spatial resolution (640x480 pixels) is unpractical for fast and thin gestures at close range (e.g. < 0.50 m). As J.Ballester and C.Pheatt concluded [16], the object size and speed requirements need to be carefully considered when designing an experiment with it. Hence, we will use the Kinect for suitable uses, aware of its limitations and capacities. Typically, the Kinect works sufficiently well between 1.40 and 3m for body tracking, with a 1cm spatial resolution at a 2m distance from an object. Furthermore, latency considerations lead us to use it for higher-level musical structures occurring in the macro space, where temporality is chosen to be loose.

Regarding the small and rapid finger gestures, we are interested in a second type of depth camera, the Leap Motion. This camera works with two monochromatic cameras and three infrared LEDs. Thanks to inverse kinematics, it provides an accurate 3D tracking of the hand skeleton, with more than 20 joints positions and velocities per hand. The Leap Motion has a lateral field of view of 150°, a vertical field of view of 120°. Its effective range extends from approximately 25 to 600mm above the camera center (the camera is oriented upwards). Additionally, the Leap Motion is known for being accurate and fast: processing time for each frame is close to 1ms, which is well below the acceptable upper bound on the computer's audible reaction to gesture fixed by [17] at 10 ms. Although additional latency will be added further with the gesture to sound mapping, the initial latency provides us with a viable starting point.

Micro and Macro bounding boxes

Igor Stravinsky (1970): *The more constraints one imposes, the more one frees one's self... the arbitrariness of the constraint serves only to obtain precision of execution.*

[3] https://www.roli.com/products/seaboard-grand

[4] http://www.eecs.qmul.ac.uk/ andrewm/touchkeys.html

We present here the design of the instrument through the 3D interactive spaces it creates. As presented before, there are three sensors: two Leap motions and one Kinect. Once placed on their slots on the EMI, the Leap Motions' field of view cover the whole surface of the table and a volume up to 30 cm above it. We designate this volume as the *micro bounding box* (figure 1). The Leap motions are centered in the halved parts of the surface while the Kinect is placed in front and above the table as displayed as displayed on figure 2. The position of the Kinect is roughly 1 – 1.20m behind the table and stands roughly 1m above the table. There is no need to place it with great precision as the system auto–calibrates the skeleton with respect to the sensors' field of view each time the software is launched.

The perspective behind the splitting of the two bounding boxes is to differentiate *macro* gestures done with the upper body with *meso/micro* gestures done with the fingers. Hence the macro space deals with the wide movements captured with the Kinect while the micro space deals with finer-grain manipulation captured with the Leap motions. Inspired by Jensenius' terminology [20], we use a unified space for both *micro* gestures happening at a millimeter scale with *meso* sound–producing gestures happening at a centimeter scale.

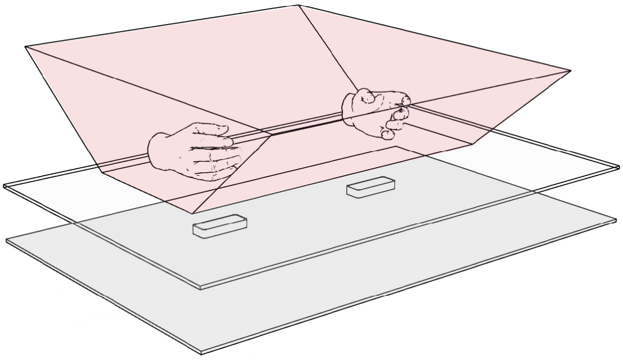


Figure 1. Micro bounding box.

MUSICAL EMBODIMENT ON A TABLETOP INSTRUMENT

One of the objective of the *Embodied Musical Instrument* is to give, through movement, meaning to the sounds thus created. If gestural electronic music performance is technically rendered possible thanks to 3D tracking devices, the coupling of perception and action, however, requires reflections on expressive use of affordance based on practice [21]. The EMI is a framework for gesture tracking and recognition, with its own metaphors and control mappings, unified within an embodied model reducing the cognitive distance between the imaginary imagery of electro-acoustic composers.

Godøy [22] distinguishes among *music imageries* images of acoustic signals, images associated with the performance, images associated with the perception and images associated

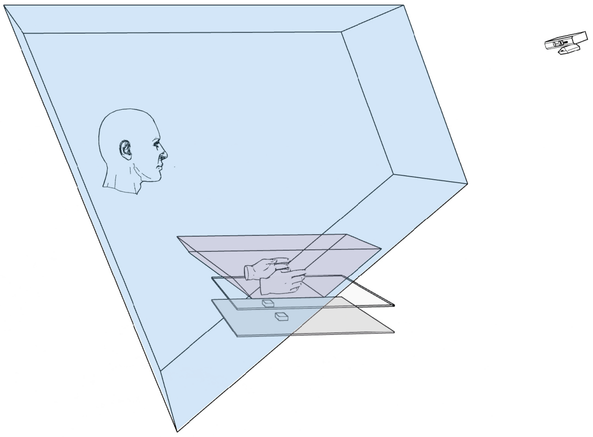


Figure 2. Macro bounding box.

with the emotive experience. He tackles the problem of understanding the nature of these sound images in our mind by drawing sketches of gestural and sonic features in a top-down manner, tracing features and sub-features of sound-morphologies and correlating them with acoustic features of sound objects. We went through a similar thought process, breaking down gestural features in *effective, accompanist and symbolic* gestures (based on Delalande' gestural typology [23]) to some lower-level gestural features enabling very specific sound controls over the articulation, intensity, aftertouch and so on. Hence, the interactive space created above the instrument can be seen as a shared gestural space including the *effective, accompanist and symbolic* gestures. In that respect, the EMI inspires an environment analogous to what Tanaka [21] and Graham[24] designate as a *performance gesture ecology*. Basically, we aim with the EMI at capturing the three categories of gestures discussed above and make use of them altogether in order to generate very expressive sounds.

The proposed metaphorical gestures we use are borrowed from keyboard instruments, touch gesture paradigms developed for touchscreen devices [11] and other physically-inspired manipulation metaphors such as an elastic cable, a wheel or a kite. The object metaphor connects to the affordance of a simple object in the mind of the user and thus, leads intuitively to the gesture to be done. We make the same assumption that the simpler the metaphor is, the more intuitive and expressive the result will be. Wessel et al. [25] similarly make this assumption as one of the necessary conditions to get an '*intimate musical control of computers*'. The other conditions being its long term potential for virtuosity, that we believe the EMI also meets with, the clarity of strategies for programming the relationship between gesture & musical results and finally a low latency response of the system.

At last, Young [26] presented how features in electro-acoustic works can be discussed through aural perception of the sound objects in association with an analytical focus based on a common understanding of the way a sound behavioral model operates. This analytical focus is however not always obvious to non initiated listeners and does not solicit the visual understandings of how things work and are produced. Physical embodiment of music performance, if realistic enough, would convey this additional information, necessary for the spectator to understand the origin of the sounds and reduce the emotional distance between the synthetic sounds and him/herself.

Metaphors in the micro bounding box

First, we were interested in building a model for dynamics, articulation and duration, inherent in the fingering. This led us to the decomposition of the fingering in several phases so as to extract information about the trajectory and the duration of each part. This representation is based on four phases: Rest, Preparation, Attack and Sustain, inspired by a more general gesture segmentation model (Preparation, Attack, Sustain, Release) [27]. Segmenting the fingering into essential phases facilitates the distinction of features for each phase (figure 3). In rest position, the hand and fingertips are relaxed on the table. In preparation, one or several fingers lift upwards. In attack, one or several fingers tap downwards while during a sustain phase, one or several fingertips stay at contact with the surface of the table. At last, the velocity of the fingertip along the z-axis in the few milliseconds prior to its contact with the table during an attack phase is mapped to sound intensity. It is worthy of note that this segmentation, which is articulated around the z-axis is only made possible thanks to the depth finger tracking of the Leap motion.

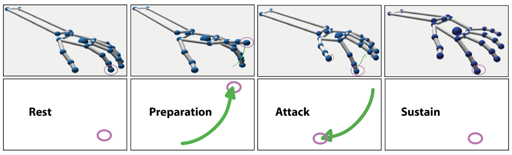


Figure 3. Rest Preparation Attack Sustain segmentation

The EMI makes use of a piano keyboard-paradigm that can be played with the fingers in the *micro–bounding box* (figure 4). The key idea here is to cover a range of notes, without the need to be extremely precise at fingering on the table since the latter is completely flat and transparent. Therefore, the zone (either blue, red or green) corresponds to a set of fives notes (e.g.: EFGAB), where each note corresponds to one finger (see colored areas on figure).

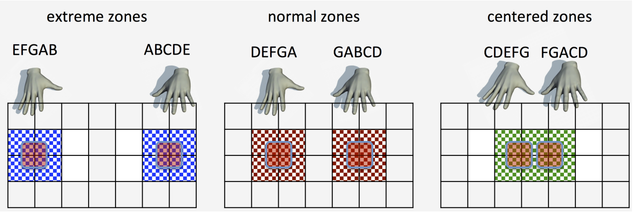


Figure 4. The keyboard paradigm.

We extend the mapping of fingertip positions on the x-axis of the table to the y-axis and attribute this dimension to the timbre space. Hence, the timbre/texture of the sound can be modified continuously by fingering at different locations along the y-axis of the table while keeping the pitch fingering system depicted above. Figure 5 depicts the top-view or avatar of a musician moving arms and hands in the pitch-timbre space.

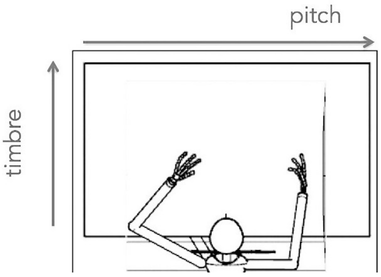


Figure 5. Pitch-timbre space.

Metaphors in the macro bounding box

From the hands' joints provided by the Kinect, we compute the three–dimensional euclidean distance between them and name this feature *elastic control*. The metaphor for lengthening/shortening the 3D euclidean between hands is to imagine that one is stretching/releasing an elastic cable. This gestural metaphor is depicted in figure 6 with the red arrow.

From the three joints *Head – Left Hand – Right Hand*, that we consider as apexes of a triangle, a plane equation is computed. Then, we respectively measure the tilt between this plane and the *xy* plane and *xz* plane of the table. The *xy* vs. *triangle* plane provides a sense of how much left or right your body is rotating, just as if one was pulling the wires of a kite or turning a wheel. Keeping on with the kite–flying metaphor, the *xz* vs. *triangle* plane reacts accordingly if the body is going backward or forward and/or the hands are going higher or lower. These two controls are represented on figure 7 respectively with the red arrow and the yellow arrow.

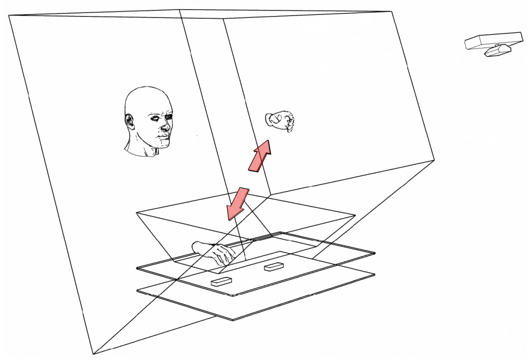


Figure 6. Elastic control: metaphor for lengthening/shortening the 3D euclidean between hands

SOUND MORPHOLOGIES MANIPULATION

In the context of a gesture-based instrument, a necessity for sound morphologies exploration is a repeatable gesture. For this matter, we use the mubu library [28] developed by the STMS team at IRCAM Sound. The mubu library, integrated into the programming language Max, embeds a movement description multi-buffer able to record gestural content in

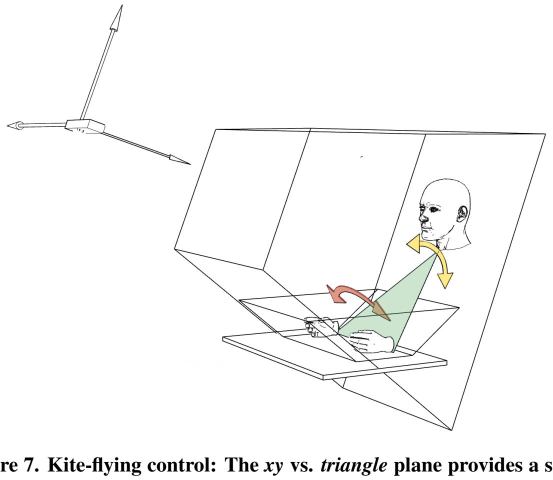


Figure 7. Kite-flying control: The *xy* vs. *triangle* plane provides a sense of how much left or right your body is rotating while the *xz* vs. *triangle* plane reacts accordingly if the body is going backward or forward and/or the hands are going higher or lower

real-time. This buffer can also be replayed, enabling to visualize the gestural data. Combined with the visual programming tool Jitter for Max, it is possible to replay the avatar of the musicians' hands and upper body. This way, one can synthesize a sound in real time with the produced gesture or replay a recorded gesture, at different speeds, forward or backward, and change the sound synthesis parameters in real time.

We discuss here how the timbre dimension is explored thanks to a physical-based sound synthesis engine named *Blotar*. It is a physical modeling synthesizer that is part of PeRColate, an open-source distribution containing a set of synthesis and signal processing algorithms for Max [29] based off the Synthesis Toolkit [30]. Physical-based sound synthesis makes sense for well articulated sounds, which we trigger when the fingers tap onto the table's surface. By changing in turn a mass, a spring or a damper parameter of the Blotar, one can oscillate between a flute and an electric guitar timbre. In our system, the brilliance parameters are mapped with the y-axis of the table's frame of reference. Hence, one can obtain brilliant sounds when the fingers tap close to the edge of the table and rounder sounds when the finger taps in the middle. Finally, the velocity of fingertips when it hits the table is mapped with the attack intensity of the sound taking into account the non-linearities occurring in such events.

Additionally, we have added a virtual piano plug-in [31] simulating physical properties and behaviors of real acoustic pianos. The EMI gesture paradigms being very much inspired by piano-like gestures, this plug-in incorporates well and despite the absence of the spring-keys haptic feedback, provides an intuitive and realistic sensation of piano playing. Finally, we have added an amplitude-convolution functionality to modify the amplitude of the Blotar sounds with the piano plug-in. Hence, one can use the attack and amplitude envelope of piano sounds with the spectral content, transients and effects of the Blotar.

Applying a new morphological frame to various spectral contents, one can reveal and enlarge some aspects such as the transients of the sounds and the sustain phase. For instance, one can imagine a noisy voice with the morphological shape of a bouncing ball or a percussive pitched sound such a vibraphone with long controllable sustain. These enables the composer to select what s/he might be interested in the sound: the shape or the content. Additionally, such physical-based and cross-synthesis techniques, already spread among composers through tools uch as Modalys [32] (IRCAM) could be handled with the EMI.

LATENCY ASSESSMENT OF THE EMI

As we are interested in evaluating the latency of the system, we are looking for the time difference between the moment when one taps onto the table and when the synthesized sound is coming out from the speakers. Therefore the experimental protocol is as follows: a microphone is plugged into a second computer, placed near by the instrument and the speakers. When the player taps on the table in order to produce a sound, the microphone picks up two signals: one is the signal produced by the physical tapping of the fingertip and the other is the synthesized sound coming from the speakers (as can be seen on figure 8). The distance between the respective attacks of the tapping sound and the synthesized sound is measured with a precision of ±2ms, as can bee seen on 9.

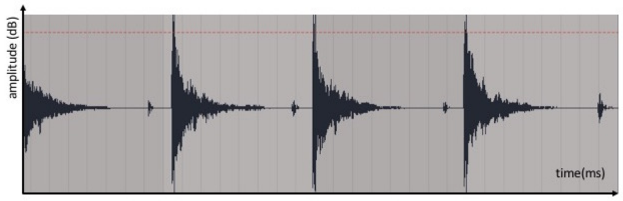


Figure 8. Recording displaying the acoustic signal of the finger tapping the acrylic sheet preceding the acoustic signal of the resulting synthesized sound

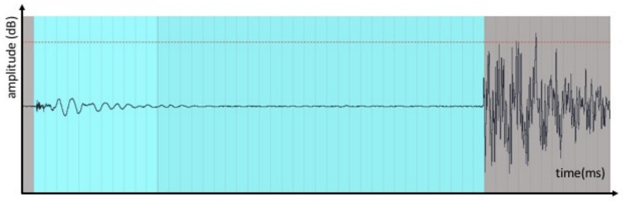


Figure 9. The blue highlighted segment corresponds to the length between the two attacks, corresponding to the latency of the system

We have recorded two sets of gestures in order to evaluate the performance of the system when one single note is repeated several times and when a sequence of notes such as an *arpeggio* is played. The two sets (1) and (2) are the following:

1. Single note repetition with index fingertip – 10 times

2. Arpeggio (thumb-middle-index-index) – 4 times

These two sets of gestures are repeated at various *beat-per-minute* (BPMs) ranging from *Lento* (60 bpm) to *Allegro* (130 bpm) and above. For each series of recording, we compute the average latency in millisecond. Additionally, we compute the *beat shift* which corresponds to unitary shift of each beat or note (eq. 1). The results are displayed on the table 1.

$$\text{beat shift} = \text{average latency} * (\text{bpm}/60) \qquad (1)$$

For instance, a beat shift equal to 0 would be a perfect temporal alignment displaying no latency at all in the system. A 0.5 beat shift musically corresponds to an off-beat and a beat shift greater than 1 occurs when the synthesized sound one hears is produced by the second previous finger tapping.

	Single Note		Arpeggio	
Tempo	Avg (ms)	Beat Shift	Avg (ms)	Beat Shift
60 bpm	154	0.15	167	0.17
70 bpm	174	0.2	184	0.21
90 bpm	230	0.34	232	0.34
110 bpm	245	0.45	273	0.5
120 bpm	304	0.6	314	0.62
130 bpm	301	0.65	345	0.74
140 bpm	369	0.86	364	0.84
160 bpm	463	1.18	419	1.12

Table 1. Average latencies and beat shifts at various BPM's for sets of gestures 1 and 2

From this table, we can see one trend: the average latency increases linearly as the BPM increases. A second observation is that the average latency is slightly greater (about 10ms) for the arpeggio than for the single note repetition.

These results are well above what is considered as acceptable for the computer's audible reaction to gesture fixed at 10 milliseconds (ms) by [17]. The Leap motion processing time per frame being 1 ms, this high audio output latency can only be explained by the typical processing scheduling delays of Max and the limit of our current OS configuration (Mid 2012 MacBook Pro Yosemite, 2.3 GHz INtel Cored i7 with 8 GB 1000 MHz DDR3 RAM). Still, it would be possible to improve the latency problem with this configuration by modifying advanced scheduling parameters in Max, such as increasing the Poll throttle, which sets the number of events processed per servicing of the MIDI scheduler and decreasing accordingly the Queue Throttle which sets the number of events processed per servicing of low-priority event queue such as graphical operations, interface events and reading files from disk. At last, it is possible to decrease the signal vector size and the sampling rate of the sound synthesis, even though this would deteriorate the overall sound quality. Further experiments will aim at finding an optimal compromise with these parameters in order to lower the latency.

CONCLUSION

In electro-acoustic music, the composer has the desire to manipulate sounds in multiple dimensions and to transform, isolate, and remix both natural and digitally created sound objects over time. One aim of the EMI is to reduce the cognitive distance between the imaginary imagery of electro-acoustic composers and the explicitly producing gestures. Embodiment seems necessary in electro-acoustic as it is intrinsic to traditional acoustical instruments and to most people approach to music. Computer music has allowed composers to use all sorts of sounds but the mechanism to produce or trigger them often do not incorporate an adequate physical movement. Realism needs a human form to physically activate processes and to avoid robotic and impenetrable performances. Novel interfaces for musical expression, such as the instrument described here, can significantly change musicians and audiences' perspectives on electronic-based music, putting forward embodied expressions through virtuoso gestures. To our knowledge, the EMI is the first musical instrument based on gesture recognition via 3D vision sensors to put forward finger expert gestures while engaging the upper body in the performance. Its ease of use is also combined with a great potential for virtuosity. The mapping strategies show transparent relationships between gestures and musical results. The latency is currently the main issue we need to solve in order to get what Wessel and Wright designate as an *intimate musical control*.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union, Seventh Framework Programme (FP7-ICT-2011-9) under grant agreement n 600676.

REFERENCES

1. Pierre Schaeffer. Traité des objets musicaux. 1966.
2. Luigi Russolo, Robert Filliou, Francesco Balilla Pratella, and Something Else Press. *The Art of Noise: futurist manifesto, 1913*. Something Else Press, 1967.
3. Gaël Tissot. La notion de morphologie sonore et le developpement des technologies en musiques electroacoustiques: Deux elements complementaires d'une unique esthetique? 2010.
4. Simon Emmerson. Computers and Live Electronic Music: Some Solutions, Many Problems. *International Computer Music Conference Proceedings*, 1991, 1991.
5. Joel Ryan. Some remarks on musical instrument design at steim. *Contemporary music review*, 6(1):3–17, 1991.
6. J Han and N Gold. Lessons Learned in Exploring the Leap Motion TM Sensor for Gesture-based Instrument Design. *Proceedings of the International Conference on New . . .* , 2014.
7. Daniel Tormoen, Florian Thalmann, and Guerino Mazzola. The Composing Hand: Musical Creation with Leap Motion and the BigBang Rubette. *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 207–212, 2014.
8. Nicolas Alessandro, Joëlle Tilmanne, Ambroise Moreau, and Antonin Puleo. AirPiano : A Multi-Touch Keyboard with Hovering Control. *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 255–258, 2015.
9. Eduardo S Silva[1], Jader Anderson O de Abreu[1], Janiel Henrique P de Almeida[1], Veronica Teichrieb, and

Geber L Ramalho. A preliminary evaluation of the leap motion sensor as controller of new digital musical instruments. 2013.

10. Chris Harrison, Hrvoje Benko, and Andrew D Wilson. Omnitouch: wearable multitouch interaction everywhere. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 441–450. ACM, 2011.

11. Craig Villamor, Dan Willis, and Luke Wroblewski. Touch gesture reference guide. *Touch Gesture Reference Guide*, 2010.

12. Sergi Jordà, Günter Geiger, Marcos Alonso, and Martin Kaltenbrunner. The reactable: exploring the synergy between live music performance and tabletop tangible interfaces. In *Proceedings of the 1st international conference on Tangible and embedded interaction*, pages 139–146. ACM, 2007.

13. Lasse Thoresen and Andreas Hedman. Spectromorphological analysis of sound objects: an adaptation of Pierre Schaeffer's typomorphology. *Organised Sound*, 12(02):129, jul 2007.

14. Otmar Hilliges. Interactions in the Air : Adding Further Depth to Interactive Tabletops. pages 139–148, 2009.

15. Hrvoje Benko and a Wilson. DepthTouch: Using depth-sensing camera to enable freehand interactions on and above the interactive surface. ... *on Tabletops and Interactive Surfaces*, (March), 2009.

16. Using the Xbox Kinect sensor for positional data acquisition. *American Journal of Physics*, 81(1):71, dec 2013.

17. Adrian Freed, Amar Chaudhary, and Brian Davila. Operating systems latency measurement and analysis for sound synthesis and processing applications. In *Proceedings of the 1997 International Computer Music Conference*, pages 479–81, 1997.

18. Wrist and shoulder posture and muscle activity during touch-screen tablet use: Effects of usage configuration, tablet type, and interacting hand. *Work*, 45(1):59–71, 2013.

19. Consumed Endurance: A metric to quantify arm fatigue of mid-air interactions. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pages 1063–1072, 2014.

20. Alexander Refsum Jensenius. Microinteraction in Music / Dance Performance. *Proceedings of the International Conference on New Interfaces for Musical Expression*, (Figure 1):16–19, 2015.

21. Atau Tanaka. Musical performance practice on sensor-based instruments. *Trends in Gestural Control of Music*, 13(389-405):284, 2000.

22. Rolf Inge Godøy. Images of Sonic Objects. *Organised Sound*, 15(01):54, mar 2010.

23. F Delalande. La gestique de gould: éléments pour une sémiologie du geste musical g. *Guertin. G. Gould, ed., Courteau, Louise*, 1988.

24. Richard Graham and Brian Bridges. Managing musical complexity with embodied metaphors. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Louisiana State University, 2015.

25. David Wessel and Matthew Wright. Problems and Prospects for Intimate Musical Control of Computers. *Computer Music Journal*, 26(3):11–22, sep 2002.

26. John Young. Sound morphology and the articulation of structure in electroacoustic music. *Organised sound*, 9(01):7–14, 2004.

27. Jules Françoise, Ianis Lallemand, Thierry Artières, Frédéric Bevilacqua, Norbert Schnell, and Diemo Schwarz. Perspectives pour l'apprentissage interactif du couplage geste-son, may 2013.

28. Norbert Schnell, Axel Röbel, Diemo Schwarz, Geoffroy Peeters, Ricardo Borghesi, et al. *MuBu and friends–Assembling tools for content based real-time interactive audio processing in Max/MSP*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2009.

29. D Trueman and R Luke DuBois. Percolate. *URL: http://music. columbia. edu/PeRColate*, 2002.

30. Perry R Cook and Gary Scavone. The synthesis toolkit (stk). In *Proceedings of the International Computer Music Conference*, pages 164–166, 1999.

31. Jukka Rauhala, Heidi-Maria Lehtonen, and Vesa Välimäki. Toward next-generation digital keyboard instruments. *Signal Processing Magazine, IEEE*, 24(2):12–20, 2007.

32. Gerhard Eckel, Francisco Iovino, and René Caussé. Sound synthesis by physical modelling with modalys. In *Proc. International Symposium on Musical Acoustics*, pages 479–482, 1995.

x2Gesture: how machines could learn expressive gesture variations of expert musicians

Christina Volioti
MTCG Lab, Department of Applied
Informatics, University of Macedonia,
156 Egnatia Street, GR-540 06,
Thessaloniki, Greece,
christina.volioti@uom.edu.gr

Sotiris Manitsaris
Centre for Robotics, MINES
ParisTech, PSL Research University,
60, Boulevard St-Michel,
75272, Paris, France
sotiris.manitsaris@mines-paristech.fr

Eleni Katsouli
MTCG, Department of Applied
Informatics, University of Macedonia,
156 Egnatia Street, GR-540 06,
Thessaloniki, Greece,
katsouli@uom.edu.gr

Athanasios Manitsaris
MTCG Lab, Department of Applied
Informatics, University of Macedonia,
156 Egnatia Street, GR-540 06,
Thessaloniki, Greece,
amanitsaris@uom.edu.gr

ABSTRACT

There is a growing interest in 'unlocking' the motor skills of expert musicians. Motivated by this need, the main objective of this paper is to present a new way of modeling expressive gesture variations in musical performance. For this purpose, the 3D gesture recognition engine 'x2Gesture' (eXpert eXpressive Gesture) has been developed, inspired by the Gesture Variation Follower, which is initially designed and developed at IRCAM in Paris and then extended at Goldsmiths College in London. x2Gesture supports both learning of musical gestures and live performing, through gesture sonification, as a unified user experience. The deeper understanding of the expressive gestural variations permits to define the confidence bounds of the expert's gestures, which are used during the decoding phase of the recognition. The first experiments show promising results in terms of recognition accuracy and temporal alignment between template and performed gesture, which leads to a better fluidity and immediacy and thus gesture sonification.

Author Keywords

expert gesture, expressive variations, musical performance, confidence bounds, gesture sonification, fluidity, immediacy

ACM Classification

H.5.2 [Information Interfaces and Presentation] User Interfaces – Interaction Styles, H.5.5 [Information Interfaces and Presentation] Sound and Music Computing.

1. INTRODUCTION

Gesture constitutes a component of human expression. It can also be characterized as a self-contained part of music. A musical performance is a sequence of expressive gestures that encapsulate both theoretical knowledge and practical motor skills. Each musical performance is unique due to expressivity, since for a given musical excerpt, interpretations can vary greatly, depending on the performer or even on expression that the performer has each time s/he plays the same piece [11].

Recently, research on the capturing and recognition of musical gestures has become very appealing. Many researchers and musicians have developed interfaces that use machine learning algorithms and aim at recognizing not only the cinematic aspects of the gesture [4][17], but also measurable parameters about expressivity [12]. From a machine learning point of view, there is usually an important compromise to make between a fast, or a rich training of the model. There are musical interfaces that are based on one-shot learning [4][12][17], in which the system requires only one training example instead of large data sets; thus, the training time is greatly reduced but significant limits are put on the modeling of expressive variations of the same gesture. Thus, the modeled information is less rich than when using large data sets. Moreover, within a sensory-motor learning context, it is important to identify precisely the tolerance between the executions of an expert performer in order to provide meaningful feedback to the learner. Therefore, the mathematical description of how an expressive gesture is being performed, along with the modeling of its variations are becoming crucial research topics.

Our approach is based on the concept that expressiveness is an intended gestural variation, which should be taken into account when modeling the gesture. In one of our previous work, Manitsaris et al. [22] has proposed a way to model offline gestural know-how in craftsmanship. As an extension of this work, we propose x2Gesture, which aims at recognizing musical expert gestures in real-time taking also into account the expressive variations. This is accomplished by implementing a) the existing work which models expert motor skills, and b) machine learning algorithms for real-time expert gesture recognition. Finally, our proposed methodology can support a unified user experience for both *learning* of expert musical gestures and *performing* musical gestures.

This paper is structured as follows: firstly, we review the state of the art (SoA) concerning machine learning algorithms that are used for gesture recognition (Section 2). Then our methodological approach (Section 3) and its implementation in two case studies (Section 4) are described. Finally, we conclude with our first evaluation results (Section 5).

2. RELATED WORK

2.1 Expert musical gestures

Firstly, we shall define some terms, which are key to our methodological approach. The term 'musical gestures' lies in the intersection between observable actions and mental

NIME'16, July 11-15, 2016, Griffith University, Brisbane, Australia.

representations [13]. A good definition of this, taken from Hatten (2003) is [19]: ‘musical gesture is biologically and culturally grounded in communicative human movement. Gesture draws upon the close interaction (and inter-modality) of a range of human perceptual and motor systems to synthesize the energetic shaping of motion through time into significant events with unique expressive force’.

When we refer to expressive gesture, what do we mean? According to [6], ‘expressiveness is conveyed by a set of temporal and spatial characteristics that operate more or less independent from the denotative meanings of those gestures’. The notion of expressivity measures *how* the expert gesture is performed. Hence, *how* an expressive gesture is performed is equally as important as *what/which* expressive gesture is performed [18].

By using the term ‘expert gestures’, we mean that performers have mastered their gestural skills. For example, they are those gestures that require years of training and practice before performers are able to perform them. Although this kind of expert has acquired high-level motor skills, expressive variations may occur between the different musical interpretations, even unconsciously. In order to control and measure expressive variations, some researchers use the ‘neutral performance’ as a reference [7], which is the performance played without any specific expressive intention. Alternatively, the mean of all the performances was taken as a reference [23].

2.2 Machine learning algorithms

Machine learning algorithms, such as those based on Hidden Markov Models (HMMs) [20], Dynamic Time Warping (DTW) [1], Hierarchical Hidden Markov Models (H-HMMs) [15], Sequential Monte Carlo technique [12] etc., are widely used for gesture recognition systems for continuous interaction. [2][3][4] successively developed a system based on a hybrid model between HMMs and DTW, called Gesture Follower (GF), for both continuous gesture recognition and following, between the template or reference gesture, and the incoming or performed gesture (template-based method). It can learn a gesture from a single example (one-shot learning), by associating each template gesture to a ‘state’ of a hidden Markov chain [5]. During the performance, a continuous estimation of parameters is calculated in real-time, by providing information for the temporal position of the performed gesture. Time alignment occurs between the template and the performed gesture, as well as offering an estimation of the time progression within the template in real-time.

One limitation of HMMs is that observations are produced at the frame level, and as a consequence they do not support the transitions between segments [15]. Therefore, [14][15] developed a system based on H-HMMs with two levels for real-time gesture segmentation and recognition. Similarly to GF, it adopts a template-based method and implements one-shot learning. The system is trained with a single pre-segmented gesture, which is annotated by the user. Each segment is associated with a high-level state (segment state), which generates the sub-models of the signal level (lower level), encoding the temporal evolution of the segment [14][16].

The aforementioned methodologies and research approaches do answer the question of what/which gesture is performed, but not how expressive gesture is performed. [12] further extended the research by proposing a template-based method which implements a Sequential Monte Carlo technique. Its main advantage is that the recognition system, named Gesture Variation Follower (GVF), is being adapted to gesture expressive variations in real-time. Specifically, in the learning phase only one example per gesture is required. Then, in the performing phase, time alignment is computed continuously and expressive variations (such as speed, size, etc.) are estimated between the template and the performed gesture [10][12].

2.3 Conclusions from SoA and Motivation

Leveraging the above, we can conclude that the majority of algorithms answer the question of what/which gesture is performed, or how it is performed, or both. Furthermore, in most cases, a parameter is implemented, measuring how much the performance is allowed to be different from template gestures [25]. Additionally, the users can control the degree of generalization of the model to ensure a robust estimation of their performed gestures with this parameter [15]. In GF and GVF, this parameter is called *tolerance* [8][25] and in [15] which is based on H-HMMs, *variance offset*. The main advantage of this parameter is that if its value is low, the system will be more robust and will recognize gestures with more accuracy. If it is set high, the system will be less reliable, due to the fact that the model will be too general and it will lead to overlaps between classes [15][25]. However, the main drawback is that the value of this parameter remains fixed during the performance of the gesture. This leads to the possibility that the system might fail to recognize some variations *within* the gesture, because it might require a slightly higher or slightly lower value of this parameter. Moreover, there is an impact on the time alignment between template gesture and performed gesture, which can vary importantly, thus reducing the immediacy and fluidity of the gesture sonification.

An additional conclusion from the literature review is that, the purpose or end-use of the implementation of algorithms is for installations, performances or even entertainment. But what happen in the case of the educational and learning process? Can the existing algorithms successfully recognize expressive gesture variations between expert and learner’s performances? For this reason, our proposed methodology deals with the know-how transmission between expert and learner. Moreover, we propose *confidence bounds*, instead of fixed values of tolerance and variance offset, which are derived from expert gesture performance [22] and can dynamically and more precisely recognize the variations that occur *within* the learner’s performance (performed gesture) in relation to the expert’s performance (template gesture). Apart from the learning the scenario, the proposed methodology gives also the possibility to the user to perform his/her own musical gestures and control sound parameters.

3. MODELING AND RECOGNITION

In the proposed methodology, the goal is not simply to train, recognize and sonify expert musical gestures, but by exploiting the existing methodologies and adding the parameter of confidence bounds, to develop a system that will be able to recognize expressive variations that take place within the gesture performance.

3.1 Expert operational model

The first step was to model expert gestural know-how in the case of the piano. This was accomplished by capturing expert musical gestures while the expert performed specific musical gestures on the piano. Then, expert gestural analysis was conducted. The purpose of using the State Space estimation methodology was two-fold: a) in order to model expert musical gestures, we built an operational model that describes how expert gestures are performed; and b) in order to develop a system that will be able to recognize more accurately the variations that might occur within the learner’s performance, we extracted the confidence bounds, based on the iterations of the same expert musical gesture, from the expert operational model [22].

The general specification of the State Space presentation of vector Y_t is given by the following dynamic system [21]:

$$\boldsymbol{Y}_t = \boldsymbol{\beta}_t + \boldsymbol{Z}_t\boldsymbol{a}_t + \boldsymbol{\varepsilon}_t \quad (1)$$
$$\boldsymbol{a}_{t+1} = \boldsymbol{\gamma}_t + \boldsymbol{W}_t\boldsymbol{a}_t + \boldsymbol{\eta}_t \quad (2)$$

where:

- $\boldsymbol{Y}_t$ is a n×1 vector, which can refer to as the signal or observation equation (1)
- $\boldsymbol{a}_t$ is an m×1 vector of possibly unobservable state variables, which can be referred to as the state or transition equation (2)
- $\boldsymbol{\beta}_t, \boldsymbol{Z}_t, \boldsymbol{\gamma}_t$ and $\boldsymbol{W}_t$ are conformable vectors and matrices
- ε_t and η_t are vectors of mean zero, Gaussian disturbances

Following equations (1) and (2), in our case the functional version of the expert operational model, presenting the gestures of the right hand with respect to dimension X (RHX), is as follows:

$$\boldsymbol{RHX}_t = \boldsymbol{Z}_{1t}\boldsymbol{a}_1 + \boldsymbol{Z}_{2t}\boldsymbol{a}_{2t} + \boldsymbol{\varepsilon}_{1t} \quad (3)$$
$$\boldsymbol{a}_{2t} = \delta_1\boldsymbol{a}_{2t-1} + \boldsymbol{\eta}_{1t} \quad (4)$$

where:

- $\boldsymbol{Z}_{1t} = [\boldsymbol{I}\ \boldsymbol{RHZ}_{t-1}\ \boldsymbol{RHY}_{t-1}\ \boldsymbol{LHX}_{t-1}],\ \boldsymbol{Z}_{2t} = [\boldsymbol{RHX}_{t-1} - \boldsymbol{RHX}_{t-2}]$
- **I** = unit vector, $\mathbf{Z}_t = [\mathbf{Z}_{1t}\ \mathbf{Z}_{2t}],\ '$ = transposition, and
- $\boldsymbol{a}_1' = [a_{10}\ a_{11}\ a_{12}\ a_{13}],\ \boldsymbol{a}_{2t}$ and δ_1 are parameters to be estimated.

Analytically, the equations to be estimated are as follows:

$$RHX_t = a_{10} + a_{11}RHZ_{t-1} + a_{12}RHY_{t-1} + a_{13}LHX_{t-1} + a_{2t}(RHX_{t-1} - RHX_{t-2}) + \varepsilon_{1t} \quad (5)$$
$$a_{2t} = \delta_1 a_{2t-1} + \eta_{1t} \quad (6)$$

In our piano case study, we mostly focused on the gestures made by playing with two hands. Thus the complete operational model has two sets of equations: three right hand equations (RHX_t, RHY_t and RHZ_t), and three left hand equations (LHX_t, LHY_t and LHZ_t).

Having estimated the system of equations (5) and (6), the expert operational model is dynamically simulated and the dependent variables are forecasted. Consequently, the estimated forecast standard error is derived according to:

$$RHX\_forecast\ se_t = s\sqrt{1 + RHX_t'(Z_t'Z_t)RHX_t} \quad (7)$$

where s = standard error of the estimated equation.

Then, we calculated the *confidence zone* for each musical gesture, including *confidence bounds* (a higher and a lower bound). The equations of the higher (8) and lower (9) bound referring to right hand are the following, where RHX_f_t is the forecasted data series at discrete time t, and RHX_se_t is the forecasted standard error:

$$RHX\_high_t = RHX\_f_t + RHX\_se_t \quad (8)$$
$$RHX\_low_t = RHX\_f_t - RHX\_se_t \quad (9)$$

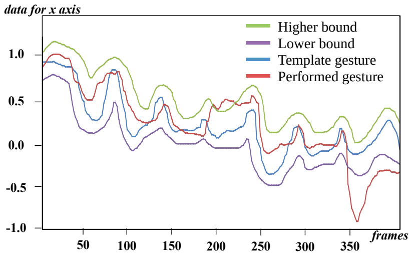


Figure 1. Confidence bounds of the expert musical gesture.

If the performed gesture is between these confidence bounds (Figure 1) during the whole performance, this means that we can successfully take into consideration the expressive variations that occur between the template and performed gesture. We further generalize this methodology by implementing the confidence bounds and using them with machine learning algorithms, in order to recognize expressive variations that might take place between the learner's (performed gesture in Figure 1) and expert's (template gesture in Figure 1) performance in real-time, in order to improve both the recognition results and the gesture sonification.

3.2 Implementation

x2Gesture is based on GVF library [1] [10][12][26], and implements the State Space and Particle Filter algorithm. The state elements are the gesture characteristics, which are for example, the time progression of the performed gesture (temporal alignment), the relative speed, the scaling coefficient (size) and the angle of rotation (orientation). The transition function is linear, relying on a Gaussian noise [9] and the observation function is the distance between the adapted template gesture and the performed gesture [10].

The algorithm includes two phases: the learning (or training) and the following (or recognition) phase. x2Gesture is first trained with a single expert example per gesture along with an audio file (pre-recorded sound). This process is repeated until the system is trained with all the template gestures, which are mapped to the respective sounds. Thereafter, in the following phase, the learner or performer imitates in real-time the same expert gesture. For each performed musical gesture, x2Gesture selects the appropriate confidence bounds, which correspond to the performed gesture. At the same time, the model aligns the incoming gesture onto the template gesture, estimating also the gesture variations [10][26]. Moreover, the system resynthesizes a plausible imitation of the original (expert) sound in real time according to the learner's gesture performance, by using the granular sound synthesis engine. The better the recognition results are, the better the gesture sonification and the re-synthesis of the sound will be.

The added value in the recognition system, as it is already mentioned, is the implementation of the confidence bounds. In this way, during the recognition, the system can prevent numerical errors that might happen due to expressive variations, and as a result, confidence bounds could improve the gesture classification and therefore the gesture sonification. This happens because confidence bounds are extracted from the expert operational model and they are not a fixed number selected by the user during the learning process or musical performance.

4. CASE STUDIES

For the evaluation of x2Gesture we organized two case studies: a) a learning scenario of expert musical gestures and b) a performance with musical gestures by using Intangible Musical Instrument (IMI) [24]. IMI setup is a construction made of Plexiglas, shaped so as to look like a table on which the learner and/or performer can put his/her hands and perform musical gestures. In both case studies, three musical gestures were included in the musical vocabulary (Table 1): a) G_1: ascending scale performed in legato style, b) G_2: descending arpeggio performed in staccato style, and c) G_3: a musical excerpt from a famous Greek song.

Table 1. (a) G_1: ascending scale, (b) G_2: descending arpeggio and (c) G_3: a musical excerpt from a Greek song

 (a)	 (b)	 (c)
Slow – 72 bps (adagio)	*Slow* – 80 bps (andante)	*Slow* – 72 bps (adagio)
Normal – 100 bps (andante)	*Normal* – 112 bps (moderato)	*Normal* – 100 bps (andante)
Fast – 116 bps (moderato)	*Fast* – 126 bps (allegro)	*Fast* – 116 bps (moderato)

[1] https://github.com/bcaramiaux/ofxGVF

All gestures have duration approximately 10-15 seconds and each user was asked to repeat each gesture five times. Apart from that, the user repeated each musical gesture in two different rhythms, slow and fast (Table 1).

In order to capture in real-time the musical gestures, two inertial sensors (Animazoo IGS-150 [2]) were used. These sensors are gyroscopes, providing XYZ axis rotations. Also they were placed on user's two hands, and specifically on wrists.

4.1 Case study I: Learning

In the learning scenario, 7 users were participated, one from whom was the expert pianist and the rest 6 were the learners. The purpose was to capture the expert pianist while performing the expert musical gestures on the piano (Figure 2 (a)). For each expert musical gesture, one iteration was selected as the reference gesture. Then, the rest iterations have been aligned and timely warped based on the reference gesture, using the DTW technique. Therefore, all the iterations of the same gesture transformed into having the same duration. These transformed data were averaged per variable and the result was used in the estimation of the expert operational model and in the extraction of confidence bounds, following the steps, which are described in Section 3.1.

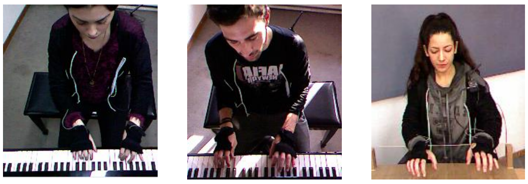

(a) (b) (c)

Figure 2. Different roles of users: (a) expert, (b) learner and (c) performer.

Subsequently, x2Gesture was trained with the three template gestures (reference). In the recognition phase, each one of the six learners performed the same expert musical gestures on the piano (Figure 2 (b)) five times. Their gestural data were captured in order to evaluate the recognition results of the model, as well as the accuracy and reliability of the confidence bounds.

4.2 Case study II: Performing

In the second case study, 6 performers were participated in total. For each performer the expert operational model and the confidence bounds were extracted. Moreover, apart from their gestural data, the sound that was produced was also recorded. Therefore, in the training phase, both reference gesture of each performers and the respective sound were given as input. In the recognition phase, each performer (Figure 2 (c)) performed the same musical gestures by using IMI, in order to resynthesize the pre-recorded sound in real-time.

5. EVALUATION

The goal of the experiment is to assess the recognition accuracy of x2Gesture, which implements the confidence bounds, comparing it also with established techniques, such as GF and GVF. The evaluation method that was used is called 'jackknife', or 'leave-one-out' approach. The basic idea is leaving out one or more observations at a time from the sample set. Practically, the database contains observations from five iterations of three musical gestures. Five distinct datasets have been created for each iteration of performed gesture. Therefore for each jackknife iteration, one dataset is left out to train the model M_i per musical gesture G_i and the rest four are used for testing. Two metrics were also used to evaluate the recognition accuracy: a) Precision, which takes into account the false recognitions and b) Recall, which takes into account the missed recognitions.

5.1 Evaluation of case study I

For the evaluation of the learning scenario, jackknife method was used only in expert's data. The aim is to evaluate the accuracy of the expert operational model and confidence bounds. Table 2 presents the results that x2Gesture gave for the five jackknife iterations, as well as the values of Precision and Recall per G_i.

Table 2. x2Gesture: Precision and Recall per expert gesture

		Maximum likelihoods			
		M_1	M_2	M_3	*Recall*
Observations	G_1	20	-	-	**100%**
	G_2	-	20	-	**100%**
	G_3	-	-	20	**100%**
	Precision	**100%**	**100%**	**100%**	

Because the results seemed to be perfect, we repeated the same experiment with GF and GVF. Therefore, Table 3 shows briefly the values of Total Precision and Total Recall per expert gesture from recognition with GF and GVF.

Table 3. GF and GVF: Precision and Recall per expert gesture

		GF		**GVF**	
		Precision	*Recall*	*Precision*	*Recall*
Observations	G_1	100%	100%	95%	100%
	G_2	100%	100%	100%	100%
	G_3	100%	100%	100%	95%
	Total	**100%**	**100%**	**98%**	**98%**

The high recognition results that x2Gesture, GF and GVF gave, can be explained by the fact that the expert pianist was very dedicated and focused on the expert performance of musical gestures. This resulted in not occurring expressive variations, even unconsciously, between the different iterations of musical interpretations. The tolerance that was used for these tests was 0.1 for both GF and GVF.

In order to complete the evaluation of the learning scenario, learners have to imitate the same expert musical gestures on the piano. The specific dataset contains: 6 learners * 3 musical gestures * 5 iterations = 90 gesture examples. The value of tolerance that was selected was 0.2 for GF and 0.1 for GVF. These tolerance values were the result of many experiments, as they gave better results for these specific musical gestures in comparison with smaller or larger tolerance values.

After having trained the system with the expert's template gesture (reference), the data from the learners' performances were given for recognition. The recognition results are presented in Table 4:

Table 4. GF, GVF and x2Gesture: expert – learners

	GF		**GVF**		**x2Gesture**	
	Precision	*Recall*	*Precision*	*Precision*	*Precision*	*Recall*
G_1	59%	57%	70%	53%	100%	70%
G_2	79%	37%	45%	43%	65%	37%
G_3	53%	83%	53%	70%	48%	83%
Total	**64%**	**59%**	**56%**	**55%**	**71%**	**63%**

According to Table 4, we can conclude that from the comparison of recognition percentages, x2Gesture gives better results than the others. These results are consistent to what we expected, and confirm the hypothesis that the recognition results can be improved with the implementation of confidence bounds. Moreover, the results confirm that confidence bounds can dynamically and more precisely recognize the variations that might occur within the learner's performance and expert's performance.

5.2 Evaluation of case study II

In the performance case study with the use of IMI, all 6 performers execute the musical gestures five times. As it is mentioned, during the

[2] http://synertial.com/

performance they were also asked to perform the gestures either slower or faster. The dataset for this case study includes per user: 3 musical gestures * 5 iterations (which contain data from slow, normal and fast speed) = 15 gesture examples. For the case study II, the value of tolerance that was selected was 0.1 for both GF and GVF.

x2Gesture was trained per user with the three template gestures along with the pre-recorded sounds. Then, in the recognition, x2Gesture selected the appropriate confidence bounds, according to the performed gestures, and resynthesized the sound in real-time, by using the granular sound synthesis engine. The recognition results per performer and per algorithm are shown in Table 5:

Table 5. GF, GVF and x2Gesture: performer – performer

		GF		GVF		x2Gesture	
		Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
User 1	G_1	75%	75%	68%	85%	64%	80%
	G_2	82%	90%	76%	65%	71%	75%
	G_3	83%	75%	61%	55%	79%	55%
	Total	**80%**	**80%**	**68%**	**68%**	**71%**	**70%**
User 2	G_1	100%	100%	100%	100%	100%	100%
	G_2	100%	100%	100%	100%	100%	100%
	G_3	100%	100%	100%	100%	100%	100%
	Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
User 3	G_1	100%	100%	95%	95%	87%	65%
	G_2	100%	100%	100%	100%	100%	90%
	G_3	100%	100%	95%	95%	67%	90%
	Total	**100%**	**100%**	**97%**	**97%**	**85%**	**82%**
User 4	G_1	71%	50%	78%	90%	95%	95%
	G_2	54%	35%	95%	90%	91%	100%
	G_3	55%	90%	89%	80%	100%	90%
	Total	**60%**	**58%**	**87%**	**87%**	**95%**	**95%**
User 5	G_1	100%	100%	95%	100%	100%	100%
	G_2	100%	100%	100%	100%	100%	100%
	G_3	100%	100%	100%	95%	100%	100%
	Total	**100%**	**100%**	**98%**	**98%**	**100%**	**100%**
User 6	G_1	100%	100%	95%	100%	100%	100%
	G_2	100%	100%	100%	95%	100%	100%
	G_3	100%	100%	100%	100%	100%	100%
	Total	**100%**	**100%**	**98%**	**98%**	**100%**	**100%**
	Grand Total	**90%**	**90%**	**91%**	**91%**	**92%**	**91%**

In the last row of Table 5, grand total from all performers are presented. If we interpret the table according to the last row, x2Gesture gives the highest results (with GVF and GF to follow).

Alternatively, if we interpret the results per performer, we can conclude that GF gives better recognition results than the others, while x2Gesture and GVF come after. This can be explained by the fact that in four out of six performers GF gives 100% in Precision and Recall, while x2Gesture in three and GVF in one. However, the majority of the percentages per performer from x2Gesture and GVF are really close to 100% (i.e. 98%, 97%, etc.), which means that the model did not manage to recognize correctly one or two gestures.

5.3 Evaluation on recognition stability and time

At this point, it is important to highlight an additional advantage of the implementation of confidence bounds. Figure 3 presents the *time progression* of the recognized G_3 from user 3 (case study II). Time index '0' is the beginning of the gesture and time index '1' is the end of the gesture. Compared to the other two algorithms, x2Gesture is more stable during the recognition process and faster than the others, because the system recognizes correctly G_3 from the 1st frame, resulting to the increase of the maximum likelihood that refer to G_3.

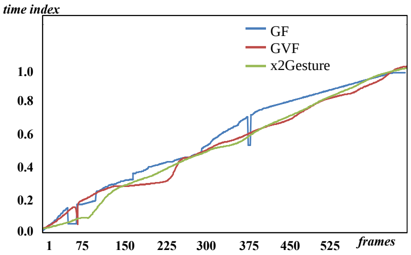


Figure 3. Gesture progression through the temporal alignments of GF, GVF and x2Gesture.

This can be also confirmed by the Figure 4(c), which presents the maximum instant likelihood. Therefore, the gesture sonification is more fluid and immediate because the new synthesized signal is much closer to the template sound. GVF becomes stable after 112 frames. Figure 4(b) shows the latency before G_3 takes the maximum likelihood. GF recognizes correctly G_3 after 145 frames, as it seems to oscillate between G_3 and G_1. The maximum likelihoods along with their transitions between gestures are presented in Figure 4(a). Although, in the end all three algorithms recognize correctly G_3, the production of the sound differs in three algorithms.

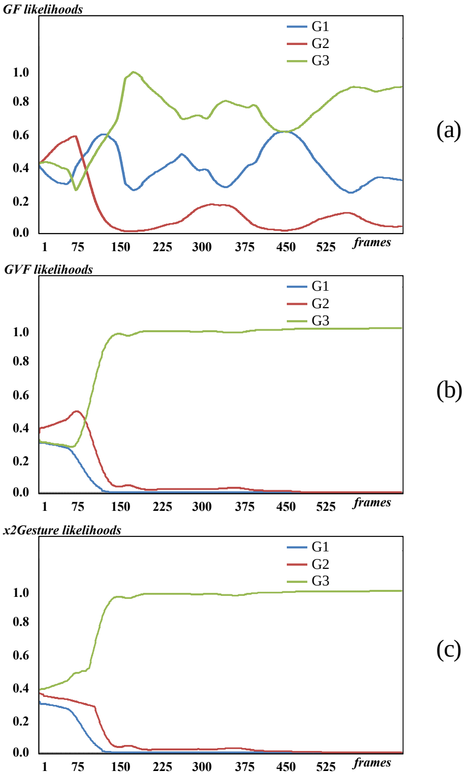


Figure 4. Instant likelihoods per frame using (a) GF, (b) GVF and (c) x2Gesture.

Additionally to the above specific example in which x2Gesture recognizes the right musical gesture faster than the other two algorithms, Table 6 presents the average time that each algorithm succeeds to recognize each musical gesture correctly.

Table 6. Average time that GF, GVF and x2Gesture need to recognize each gesture correctly

	GF		GVF		x2Gesture	
	Mean	*St.Dev.*	*Mean*	*St.Dev.*	*Mean*	*St.Dev.*
G_1	8,43	7,27	2,25	1,65	2,23	1,53
G_2	1,94	3,17	3,25	2,56	1,65	1,73
G_3	0,71	1,34	2,46	2,43	0,96	0,68

In order to evaluate the response time in real time, the database from case study I (expert – learners), was used. According to the small values of mean and standard deviation for each gesture, it can be further confirmed that x2Gesture can recognize faster and more stable the musical gestures without oscillating between all three musical gestures. However in G_3, GF has smaller mean value than x2Gesture, but larger standard deviation. This can be interpreted by the fact that, although GF has recognized more G_3 in comparison to x2Gesture (Precision in Table 4), the values of time that GF has taken the highest instant likelihoods for G_3 varied with each other more (max. time value 5,02 sec. and min. time value 0,11 sec.) than in x2Gesture (max. time value 3,01 sec. and min. time value 0,13 sec.).

6. CONCLUSION AND PERSPECTIVES

Summarizing, we propose the 3D gesture recognition engine 'x2Gesture', which has been especially designed to address the needs of both learning the expert musical gestures and live performing through gesture sonification. Moreover, the proposed modeling of the expressive variations and the output confidence bounds, led to higher recognition accuracy even in multi-user use-cases, by taking into consideration the expressive variations that might occur. Furthermore, the first evaluation results prove that there is a more fluid and immediate temporal alignment with the correct gesture.

Our future work is to generalize our methodology in order to be used in a variety of different disciplines, by creating connections between them. For example, to combine music with mathematics, or physics, or drawing, etc. Expressivity and creativity will be the core of these interdisciplinary musical performances.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union, Seventh Framework Programme (FP7-ICT-2011-9) under grant agreement n° 600676.

We wish to express our sincere appreciation to all musicians who took part in this experiment. Also, we would like to thank Gavriela Senteri for her contribution in the validation of the models.

8. REFERENCES

[1] F. Bettens, and T. Todoroff. Real-time dtw-based gesture recognition external object for max/msp and puredata. *In Proc. of the SMC 2009 Conference*, 30, 35, 2009.

[2] F. Bevilacqua, F. Guédy, N. Schnell, E. Fléty, and N. Leroy. Wireless sensor interface and gesture-follower for music pedagogy. *In Proc. of the NIME'07*, New York, 2007, 124-129.

[3] F. Bevilacqua, R. Muller, and N. Schnell. MnM: a Max/MSP mapping toolbox. *In Proc. of the NIME'05*, Vancouver, Canada, 2005.

[4] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. *In Proc. of the 8th International Conference on Gesture in Embodied Communication and Human-Computer Interaction,* Bielefeld, Germany, 2009.

[5] A.F. Bobick, and A.D. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE TPAMI*, 19, 12, 1997, 1325-1337.

[6] A. Camurri, G. De Poli, M. Leman, and G. Volpe. A multi-layered conceptual framework for expressive gesture applications. *In Proc. of the International MOSART Workshop*, Barcelona, Spain, 2001.

[7] S. Canazza, G. De Poli, C. Drioli, A. Roda, and A. Vidolin. Modeling and control of expressiveness in music performance. *In Proc. of the IEEE*, 92, 4, 2004, 286-701.

[8] B. Caramiaux. Études sur la relation geste–son en performance musicale. *Ph.D. Thesis*, Pierre and Marie Curie University (Paris 6), France, 2011.

[9] B. Caramiaux. Motion Modeling for Expressive Interaction: A Design Proposal using Bayesian Adaptive Systems. *In Proc. of the MOCO'14*, Paris, France, 2014.

[10] B. Caramiaux. Optimising the Unexpected: Computational Design Approach in Expressive Gestural Interaction. *In Proc. of the CHI Workshop on Principles, Techniques and Perspectives on Optimization and HCI*, Seoul,Korea,2015.

[11] B. Caramiaux, M. Donnarumma, and A. Tanaka. Understanding Gesture Expressivity through Muscle Sensing. *ACM Trans. on Computer-Human Interaction*, 2, 6, 2015.

[12] B. Caramiaux, N. Montecchio, A. Tanaka, and F. Bevilacqua. Adaptive Gesture Recognition with Variation Estimation for Interactive Systems. *ACM TiiS*, 4, 4, 2015.

[13] F. Delalande. *La gestique de gould: Élements pour une sémiologie du geste musical.* In: G. Guertin (Eds.), Glenn Gould Pluriel, Québec: Louise Courteau, 1988, 85–111.

[14] J. Françoise. Gesture-Sound Mapping by Demonstration in Interactive Music Systems. *In Proc. of the 21st ACM MM'13,* Barcelona, Spain, 2013, 1051-1054.

[15] J. Françoise. Motion-sound mapping by demonstration. *Ph.D. Thesis*, Pierre and Marie Curie University, France, 2015.

[16] J. Françoise, B. Caramiaux, and F. Bevilacqua. A Hierarchical Approach for the Design of Gesture-to-Sound Mappings. *In Proc. of the CMC Conference*, Copenhagen, Denmark. 2012.

[17] J. Françoise, N. Schnell, R. Borghesi, and F. Bevilacqua. Probabilistic Models for Designing Motion and Sound Relationships. *In Proc. of the NIME'14*, London, UK, 2014.

[18] N. Gillian. Gesture Recognition for Musician Computer Interaction. *Ph.D. Thesis*, Queen's University Belfast, UK, 2011.

[19] R.S. Hatten. Musical Gesture: Theory and Interpretation. *Course note*s, Indiana University, 2003, http://www.indiana.edu/~deanfac/blfal03/mus/mus_t561_9824.html (Accessed 5 January 2016).

[20] P. Kolesnik, and M.M. Wanderley. Implementation of the Discrete Hidden Markov Model in Max/MSP Environment. *In Proc. of the FLAIRS*, 2005, 68-73.

[21] S. J. Koopman, N. Shephard, and J. A. Doornik. Statistical algorithms for models in state space using SelfPack 2.2. *Econometrics Journal*, 1, 1998, 1-55.

[22] S. Manitsaris, A. Glushkova, E. Katsouli, A. Manitsaris, and C. Volioti. Modelling Gestural Know-how in Pottery Based on State-space Estimation and System Dynamic Simulation. *Procedia Manufacturing*, 3, 2015, 3804-3811.

[23] B.H. Repp. Diversity and commonality in music performance: an analysis of timing microstructure in schumann's "traumerei". *Journal of the Acoustical Society of America*, 92, 1992, 2546-2568.

[24] C. Volioti, E. Hemery, S. Manitsaris, V. Tsekouropoulou, E. Yilmaz, F. Moutarde, and A. Manitsaris. Music Gestural Skills Development Engaging Teachers, Learners and Expert Performers. *Procedia Manufacturing*, 3, 1543-15, 2015.

[25] B. Zamborlin, F. Bevilacqua, M. Gillies, and M. D'inverno. Fluid gesture interaction design: Applications of continuous recognition for the design of modern gestural interfaces. *ACM TiiS*, 3, 4, 2014, 1-30.

[26] A.V. Zandt-Escobar, B. Caramiaux, and A. Tanaka. PiaF: A Tool for Augmented Piano Performance Using Gesture Variation Following. *In Proc. of the NIME'14*, London, UK, 2014.

A Natural User Interface for Gestural Expression and Emotional Elicitation to access the Musical Intangible Cultural Heritage

CHRISTINA VOLIOTI[1], University of Macedonia
SOTIRIS MANITSARIS[2], MINES ParisTech
EDGAR HEMERY[2], MINES ParisTech
STELIOS HADJIDIMITRIOU[3], Aristotle University of Thessaloniki
VASILEIOS CHARISIS[3], Aristotle University of Thessaloniki
LEONTIOS HADJILEONTIADIS[3,4], Aristotle University of Thessaloniki
ELENI KATSOULI[1], University of Macedonia
FABIEN MOUTARDE[2], MINES ParisTech
ATHANASIOS MANITSARIS[1], University of Macedonia

This paper describes a prototype natural user interface, named the Intangible Musical Instrument, which aims to facilitate access to the knowledge of the performers that constitutes musical Intangible Cultural Heritage, using off-the-shelf motion capturing that is easily accessed by the public at large. This prototype is able to capture, model and recognize musical gestures (upper body including fingers) as well as to sonify them. The emotional status of the performer affects the sound parameters at the synthesis level. Intangible Musical Instrument is able to support both learning and performing/composing by providing to the user not only intuitive gesture control but also a unique user experience. In addition, the first evaluation of the Intangible Musical Instrument is presented, in which all the functionalities of the system are assessed. Overall, the results with respect to this evaluation were very promising.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning – Knowledge acquisition; H.m [**Information Interfaces and presentation (e.g., HCI)**]: Miscallaneous.

General Terms: Intangible Cultural Heritage, Musical gestures, Expert knowledge, Natural User Interface

Additional Key Words and Phrases: Gesture Recognition, Emotional status, Sonification, Evaluation

1. INTRODUCTION

Cultural expression is not limited to architecture, monuments or collections of artifacts. It also includes fragile intangible live expressions, which involve knowledge and skills. Such expressions include music, human skills, etc.. These manifestations of human intelligence and creativeness constitute the Intangible Cultural Heritage (ICH)[1]. ICH is at the same time traditional, contemporary

This work is supported by the Widget Corporation Grant #312-001.
Author's address: C.Volioti, E-mail: christina.volioti@uom.edu.gr. S.Manitsaris, E-mail: sotiris.manitsaris@mines-paristech.fr. E.Hemery, E-mail: edgar.hemery@mines-paristech.fr. S.Hadjidimitriou, E-mail: stelios@psyche.ee.auth.gr. V.Charisis, E-mail: vcharisis@ee.auth.gr. L.Hadjileontiadis, E-mail: leontios@auth.gr. E.Katsouli, E-mail: katsouli@uom.edu.gr. F.Moutarde, E-mail: fabien.moutarde@mines-paristech.fr. A.Manitsaris, E-mail: amanitsaris@uom.edu.gr. [1]Multimedia, Security and Networking Lab, Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece. [2]Centre for Robotics, MINES ParisTech, PSL Research University, Paris, France. [3]Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece. [4]Department of Electrical and Computer Engineering, Khalifa University of Science and Technology, Abu Dhabi, UAE.

https://doi.org/10.1145/3127324

[1] http://i-treasures.eu/

and living, because it does not only refer to inherited knowledge but also to the renewal of contemporary cultural expressions. It refers to the past, to the present, and, certainly to the future and is the mainspring of humanity's cultural diversity. According to UNESCO, music is the most universal form in the performing arts, since it can be found in every society, usually as an integral part of other performing art forms and other domains of the ICH. Music of different types such as classical, contemporary or popular, sacred etc., can be found in a large variety of contexts. Instruments, artefacts and objects, in general, are closely linked with musical expressions and they are all included in the Convention's definition of the ICH [UNESCO 2003]. Music that fits with the Western form of musical notation is better protected, while those that do not fit are usually threatened with disappearance when their holders die. Thus, the crucial point for all music forms is to develop the motor skills of playing a musical instrument by strengthening the bond between the expert performer and the learner. Motivated by this need, in recent years, researchers focused on the study of embodiment and enactive concepts. These concepts reflect the contribution of body movement to the action/perception and the mind/environment interaction [Noë 2004]. In the performing arts, and, more precisely, in music, body movement is semantically connected with gesture in most activities, such as performing and composing.

Composers bring together knowledge and skills in sound coloring and organization, in terms of structure and form. These skills are depicted on the music score of their pieces, which constitute the Tangible Cultural Heritage (TCH). Nevertheless, the music score usually contains only a few abstract annotations about idiomatic gestures that should be incorporated by the performer during his/her musical and physical playing. Such information leads to the organization of the musical material, which is culminated in a compositional structure. The analysis of the musical material always brings to the surface the question "how does this work?". Music theory explains the musical structure and/or defines the way the material functions, according to various viewpoints, such as those of Allen Forte, Arnold Whittall, Rosemary Killiam and Patrick McCreless [Hadjileontiadis 2014]. Therefore, music theory can explain how a piece of music functions, but it does not provide information about the method and more precisely, the way the musician should interpret the musical score or/and how to perform.

The performance is the result of the symbiosis between the musician and his/her instrument. This symbiosis takes the form of an interactional relationship, where the musician is both a trigger and a transmitter, connecting the *perception* (mediated instrumental mechanisms and physical environment), the *knowledge* (theoretical understanding of the inherited music score) and the *gesture* (semantic motor skills). Consequently, the expert musical gesture can be considered as a fully embodied notion that encapsulates the motor skills of the performer to interpret musical pieces, following the musical notation defined by the composer. Moreover, the musical instrument is a physical interface that can be considered as a means of musical expression and performance. Nevertheless, the learning curve of playing musical instruments requires years of training, practice, and apprenticeship before being able to perform. Furthermore, the learning of expert musical gestures is still viewed as a communicative act of social interaction, rather than "my own" personal experience. Consequently, "learning" musical gestures and "performing" music are usually perceived as separate concepts and experiences. This means that accessing knowledge is a long-term procedure, since there is no quick transition from novice to expert.

Based on the above need, the purpose of this paper is to present a Natural User Interface (NUI) named the "Intangible Musical Instrument" (IMI) for capturing, modeling and recognition of musical gestures of expert performers which will be able to support both "learning" and "performing/ composing" as a unified user experience.

2. STATE OF THE ART

2.1 Musical gestures as referential patterns of composers

Gesture is the core activity of music creation; a dynamic organism, similar to the human organism; an experience that combines structural properties of music together with cultural and historical contexts [Truslit 1938; Coker 1972; Broeckx 1981; Hatten 1994; Cadoz and Wanderley 2000; Cumming 2000]. In talking about musical gestures and cultural heritage, there is an endless list of composers and knowledge that constitute an ICH. For example, short musical patterns, which can easily be imitated through body gestures, constitute, for Beethoven, the palette of his compositions. These short patterns, and their variations, constitute an ongoing unfolding process throughout his musical pieces. Many analysts consider this practice as a self-referential context where musical gestures, similar to other variations of the same gesture, are recognized within the same piece. Another example, which can be given is the musical collage of gestures in the Sinfonia of Berio. Gestural patterns of Mahler, Ravel and Debussy are integrated into the new musical piece so that in the Sinfonia they remain as representative musical idioms that transmit music-related cultural meanings [Godøy and Leman 2009].

Consequently, "musical patterns" and "gestural patterns in music" are closely linked notions, since sonic forms are understood through embodiment. These patterns constitute elements of social interaction and differentiation since their imitation entails the acquisition of cultural models for emulation. According to McNeill [1992], these patterns can be considered throughout history as playing an important role in creating and sustaining human communities and can be understood as a mirror system between composer and listener or even master and learner [Clayton 2000; Keller 2008]. Nevertheless, musical pieces documented through musical scores, which constitute a TCH, encapsulate only abstract information about energy and expressivity of gestures, which are finally incarnated through the interpretation of performers on musical instruments.

2.1.1 Typology of musical gestures

The gesture vocabulary described by Delalande, which has been extensively used in the literature [Delalande 1988; Cadoz and Wanderley 2000; Zhao 2001], divides musical gestures into three classes. The first class is named "*effective gesture*" and it concerns movements that are necessary to mechanically produce the sound (e.g. press a key). The second class is named "*accompanist gesture*" and it refers to sound-facilitating movements (e.g. specific postures that permit expressivity). Finally, the "*figurative gesture*" conveys symbolic messages to the audience as a communication act.

2.1.2 Transmission of gestural know-how in music

The examples documented in the previous two sections show that the musical meaning of gestural know-how involves different levels of information, which are: a) first-person, b) second-person and c) third-person perspectives on gesture [Leman 2010].

The *first-person perspective* on gesture defines the meaning of the gesture for the person that actually implements it. Within the ICH context, the expert performers are holders of ICH that have perfected their know-how to include high-level specific characteristics. Additionally, the learner can also have a first-person perspective when playing a musical instrument. The difference between the two is that the expert has developed, at a greater level than the learner (it really depends on the level of the learner), his/her action-based approach to gesture, because s/he knows all the gestural patterns in music. S/he has mental access to how the action, described on the musical score, is deployed over time and s/he has the capacity to control his/her sensorimotor system that produces the corresponding sonic form.

The *second-person perspective* on gesture refers to how other people perceive the musical gesture in a social interaction context. This approach is the most typical one that is used in music schools, conservatories, etc.. The learner observes the experts, which in most cases are his/her teacher, following the concept of “my” perception of “your” gesture [Leman 2010]. According to this “me-to-you” relationship, a mirroring system is established between expert and learner, where the body movements of the learner are deployed, so that the movement of the expert, incorporating the knowledge of the composer derived from the musical score, is understood as an action by the learner.

The *third-person perspective* on gesture focuses on the measurement and capturing of moving objects. This task can be done by a computer using audio recording, video recording, motion capture technologies and brain scans, as well as physiological body changes [Pratt 1931/1968; Friberg and Sundberg 1999; Camurri et al. 2005]. In this way, the knowledge of the performer is captured, based on techniques of feature extraction and pattern matching.

2.2 Emotional expression in music

There is no need for scientific evidence to support the fact that music expresses emotions, as personal accounts of affective experiences during listening to music are more than sufficient. However, a vast amount of research has been conducted in order to reveal further insights into this phenomenon, ranging from philosophical to biological approaches [Juslin and Sloboda 2010]. It has been suggested that such music-induced emotions are governed by universality in terms of musical culture, meaning etc. and that listeners with different cultural backgrounds can infer emotions in culture-specific music to a certain extent. Such evidence has led to the assumption that neurobiological functions underlying such emotional experiences do not differ across members of different cultures, as the responsible neural networks may be fixed. In general, the processing of musical stimuli involves the gradual analysis of music structural elements from basic acoustic features to musical syntax that leads to the perception of emotions and semantic meanings underlying the stimuli [Koelsch and Siebel 2005]. It is becoming evident that the structure of music defines what it expresses. To be more accurate, music does not literally express emotion, but it is its structural elements and production performance shaping the acoustic outcome that foster the induction of emotional states in the listener. In the following descriptions, affective states will often be characterized based on the *valence-arousal model* [Russell 1980]. Valence denotes whether an emotion is positive or negative, while arousal refers to the level of excitation that the emotion encapsulates.

2.2.1 Emotions in musical performance

Written music can be performed in different ways just as a piece of text can be read with various tones. In an important sense, it can be argued that music and performances of the same work can differ significantly. The latter form the concept of performance expression that refers to both a) the correlation between the performer's interpretation of a musical excerpt and the small-scale variations in timing, dynamics, vibrato, and articulation that shape the microstructure of the performance and b) the relationship between such variations and the listener's perception of the performance. It has been proposed that performance expression emerges from five different sources, i.e. Generative rules, Emotional expression, Random fluctuations, Motion principles, and Stylistic unexpectedness, referred to as the GERMS model [Juslin 2003]. Here, the focus is placed on emotional expression that allows the performer to convey emotions to listeners by manipulating features such as tempo and loudness in order to render the performance with the emotional characteristics that seem suitable for the particular musical piece. Table I reports the primary acoustic cues of emotional expression in music performance [Gabrielsson and Lindström 2010; Juslin and Timmers 2010]; these are mainly empirical relationships, rather than absolute, and constitute an appealing research topic.

Table I. Empirical relationships between sound parameters and emotions [Gabrielsson and Lindström 2010; Juslin and Timmers 2010]

PARAMETERS	DEFINITION	ASSOCIATED EMOTIONS
Tempo	The speed or pace of a musical piece	Fast tempo: happiness, excitement, anger Slow tempo: sadness, serenity

Mode	The type of scale in which the piece is written	Major tonality: happiness, joy Minor tonality: sadness
Loudness/ Volume	The physical strength and amplitude of a sound	Loud sound: happiness or power, anger Soft sound: relaxation, tenderness or sadness
Melody	The linear succession of musical tones that the listener perceives as a single entity	Complementing harmonies: happiness, relaxation Clashing harmonies: excitement, anger
Tonality	Musical key or the relations between the notes of a scale or key of a musical piece.	Tonal: joyful, dull Atonal: angry
Rhythm	The regularly recurring pattern or beat of a song	Smooth/consistent rhythm: happiness, peace Rough/irregular rhythm: amusement, uneasiness

It is clearly conceivable that emotions play a significant role in musical artistic expression. Consequently, the analysis and manipulation of users' affective states should be taken into serious consideration within Intangible Musical Instrument (IMI) design, development and practice that aims to support music performance.

2.3 Gesture control of sound

In order for a musical interface, or instrument, which draws gestural data from sensors and cameras, to feel natural from the point of view of user experience, it should provide intuitive gesture control of sound. With the term "mapping gesture to sound" or "gesture sonification" is meant the procedure, in which the gestural data is being associated with the sound parameters; therefore, the gesture characteristics and features, as well as the sound synthesis variables that are going to be used, have to be defined. Then, a decision about the strategy of mapping, explicit or implicit mapping, also has to be made. In *explicit* mapping, also called direct mapping, the input is directly associated to the output while, *implicit* or indirect mapping refers mostly to the use of machine learning techniques, which imply a training phase to set parameters [Bevilacqua et al. 2011].

2.4 Conclusions from the state-of-the-art and motivation

Leveraging the above, there is a growing interest in the analysis of the gestural knowledge. A large amount of studies conducted in the last years on embodied music cognition have investigate that not only effective, but also accompanist and figurative gestures are very important, since they are related to the expressivity and to the social interaction of the performer with the audience [Jensenius 2007; Maes et al. 2010]. However, "learning" musical gestures and "performing" music are usually perceived as separate concepts and experiences that pass through intermediate physical mechanisms. Usually, for learners, the "challenges" related to the physical aspects, such as the instrument being more important than the "skills" needed in music playing, can cause frustration. Consequently, the achievement of good motor skills is a long-term procedure. Additionally, the learning of the motor skills that are self-referential for a specific performer still constitute a "black box" for the learner, since it can be only approached as a second-person experience; therefore, when the learner observes the expert, s/he perceives as expert motor skills, the limited abstract sonic movements, which are visually derived from the expert gestures. Finally, the skills required of the performer, whether gestural or emotional, are not documented on the music score in much detail. Hence, in cases where a musical piece does not follow the Western form of music notation, it is extremely difficult to transmit it to the next generations.

Motivated by the above, the main objective of the present work is to create a natural user interface of the gestural expression and emotion elicitation in music. This natural user interface refers to Intangible Musical Instrument (IMI), which provides a holistic approach to gesture capturing, recognition and sonification, taking into consideration the emotional status of the performer at the synthesis level. Moreover, IMI can support learning, performing and composing with gestures as a first-person experience, by putting the user at the core of musical activities, such as performing and composing with gestures.

3. METHODOLOGICAL APPROACH

3.1 Methodology of gestures and emotions sonification

IMI supports the continuous and real-time gesture control of sound, taking into consideration the emotional status of the performer. Therefore, the fundamental elements of the proposed methodology (Figure 1) are the modalities that are involved in the musical performance, which are the gestures, the emotions and the sound. The scientific challenge of this methodology is to propose a coherent way of interconnecting these modalities, and thus answering "what to map, where and how?".

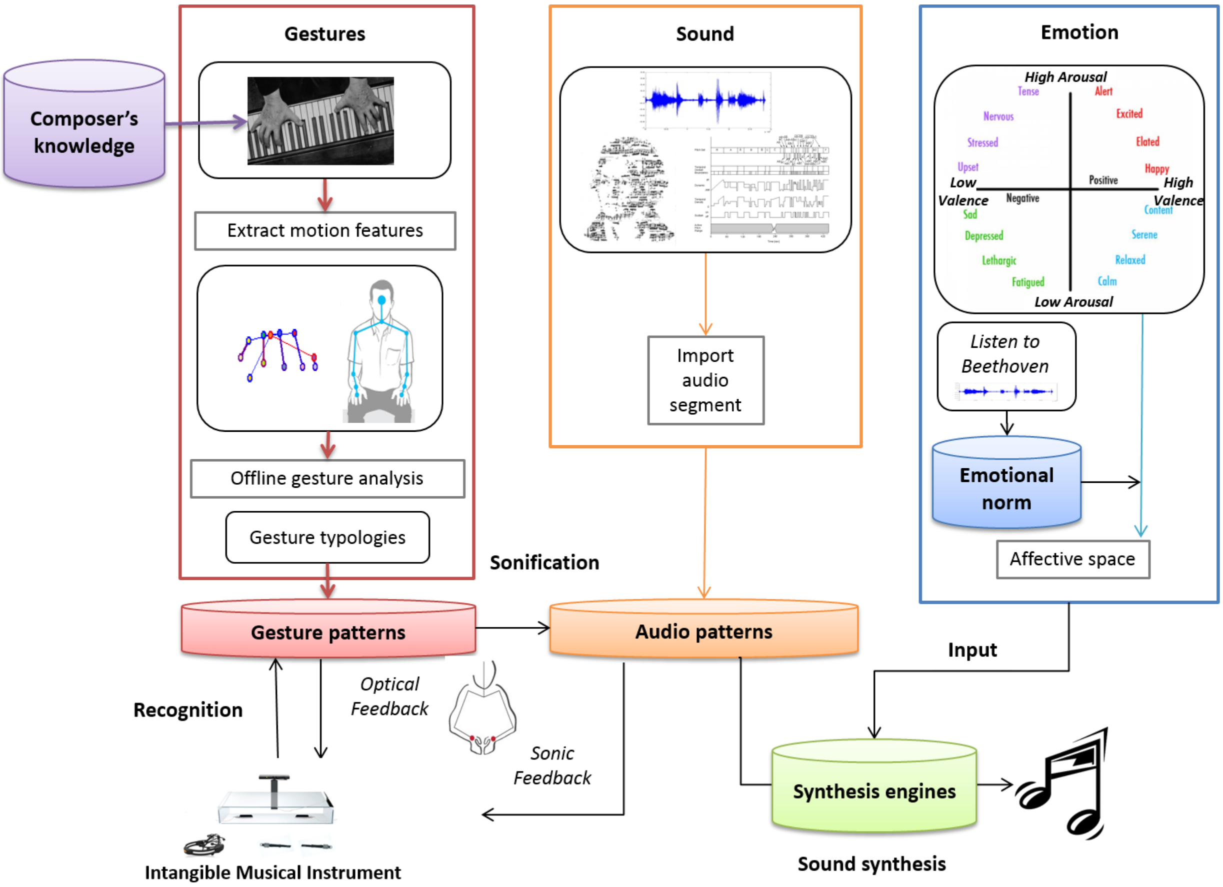


Fig. 1. Methodology of gesture sonification taking into consideration the emotional status of the performer

The motor skills of the user are put at the core of the sound creation, by using the gestural modality as a trigger of sound processes. These motor skills are captured by motion capture sensors, which are exactly the same for all the user profiles (e.g. expert, learner) for both learning and performing/composing. Taking into consideration the need for modeling the expert gestures, robust gestural information is captured in order to apply on it machine learning and pattern recognition methods. All the movements of the upper body part, including finger and head motions, are captured and represented through measurable physical descriptors; therefore, the physiological analysis of gestures focuses on an analytical description based on cinematic, spatial and frequential characteristics. More precisely, a hybrid rotational and Cartesian representation of the motion is applied, using inertial sensors and depth cameras. The descriptors of the expert gestures are used to create deterministic models (on a frame-by-frame basis) as well as to train stochastic models based on time series. In a learning context, when the learner performs a gesture, his/her gestural descriptors are compared online with the expert models of all the gestures of the vocabulary and a gesture is recognized according to which model outputs the highest probability. In a musical performance context, the performer can train the models with his/her own gestural data and recognize them online. The gesture recognition engine that uses time series as input is based on a hybrid approach of Hidden Markov Models (HMMs) and Dynamic Time Warping (DTW) [Bevilacqua et al. 2007], where HMMs are used to recognize the gesture and DTW to temporarily align the modeled gesture with the input gesture.

As long as the gesture is recognized, different mapping strategies are proposed. The first strategy refers to the connection of gesture perceptual parameters to some set of sound perceptual parameters, which are translated into concepts that can be perceived visually (gestures) or sonically (sounds). This strategy is also known as explicit mapping and it is used for associating fingerings to pitches as a one-to-one relationship. Similarly, with bijective functions, there is a mapping between 3D positions of fingertips to the creation of specific notes. This function takes gesture as input and outputs sounds with a MIDI piano synthesizer. When the fingertips come into contact with the surface of the IMI or hover less than a centimeter from it, a note is produced, the sound of which is determined by a set of parameters such as speed and the fingers' trajectory before contact. Moreover, the musically interactive surface is articulated in three zones, similar to the octaves of acoustic pianos, and each of them is associated with the hand's centroid.

The second mapping strategy, called implicit mapping, is based on a temporal mapping method [Bevilacqua et al. 2011]. The basic advantage of this approach is the time warping of the sound that is produced, depending on the speed of the performed gesture in real-time. It replays sound samples at various speeds, according to the gesture performed in real-time. Audio time stretching and compressing, as well as re-synthesis of audio can be accomplished by using the granular sound synthesis engine. In particular, the temporal mapping method associates a sound with a template gesture and links temporal states of a sound with the temporal states of the template gesture. Implicit mapping is based on information that is given from head, arms and the vertebral axis, meaning the upper body, without including the fingers.

Finally, music is well-known for affecting human emotional status, but the relationship between specific musical parameters and emotional responses is still not clear. Taking into account Table I, the sound parameters that are proposed in this research and are directly associated to the emotional status (Valence-Arousal model) are the loudness and the pitch. More specifically, the values of Valence modify the pitch of the sound, while the values of Arousal change the loudness. In both learning and performing/composing contexts, the Valence and Arousal parameters of the user are used as input to the sound synthesis engine, thus, mostly affecting the intensity and the timbre of the sound.

3.1.1 Learning the expert musical gestures

As has already been mentioned, the key point for all music forms is to have access to the gestural knowledge of playing a musical instrument and the strengthening of the bond between the expert holder of the ICH (which is the composer or performer) and the learner. As a result, in a learning scenario, the learner performs pre-defined expert gestures, taken from the vocabulary. Therefore, s/he imitates these expert gestures. S/he attempts to get close enough to the expert gesture model, so that the sound can be re-synthesized at its original speed. The re-synthesis of the sound is based on the granular sound synthesis engine.

3.1.2 Composing with gestures

In a composing scenario, the composer has the ability to create his/her own vocabulary of musical gestures, to describe the expressiveness by defining the appropriate emotions that the performer should imitate and to sonify his/her gestures and emotions by defining the sonic spaces and parameters. As a result, the composer is able to experiment with his/her own gesture-sound mappings and audio synthesis, as well as to compose contemporary music by performing gestures one after the other, by using fingers, body gestures and emotions. The goal of the composing scenario is to provide a generic system, which can be adapted according to the needs of each performer and composer. The variety of sounds the IMI can produce is equivalent to most synthesizers, but the way the musician interacts with it is totally unique, making the interface a powerful tool for both performing and composing music. However, it is important to highlight that the IMI is not a virtual replacement for the piano (or any other keyboard instrument), but an adaptation of the existing techniques for this instrument to computer music, including electronic and electroacoustic music.

4. TECHNICAL IMPLEMENTATION AND SOFTWARE DEVELOPMENT

The aforementioned methodology, which refers to capturing, analyzing, recognizing data, mapping gesture to sound, as well as sound synthesis, is implemented with Max/MSP programming language[2]. The setup prototype is a construction made of Plexiglas, shaped so as to look like a table on which user can put his/her hands (Figure 2). The dimensions of the table are 70 cm long, 40 cm wide and 13 cm high. The setup lies on a table so that the hands on the Plexiglas are placed at a comfortable height.

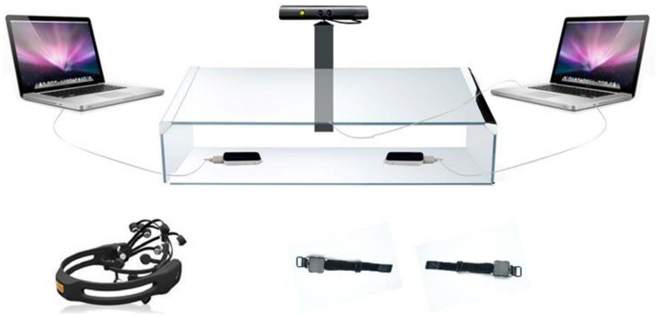
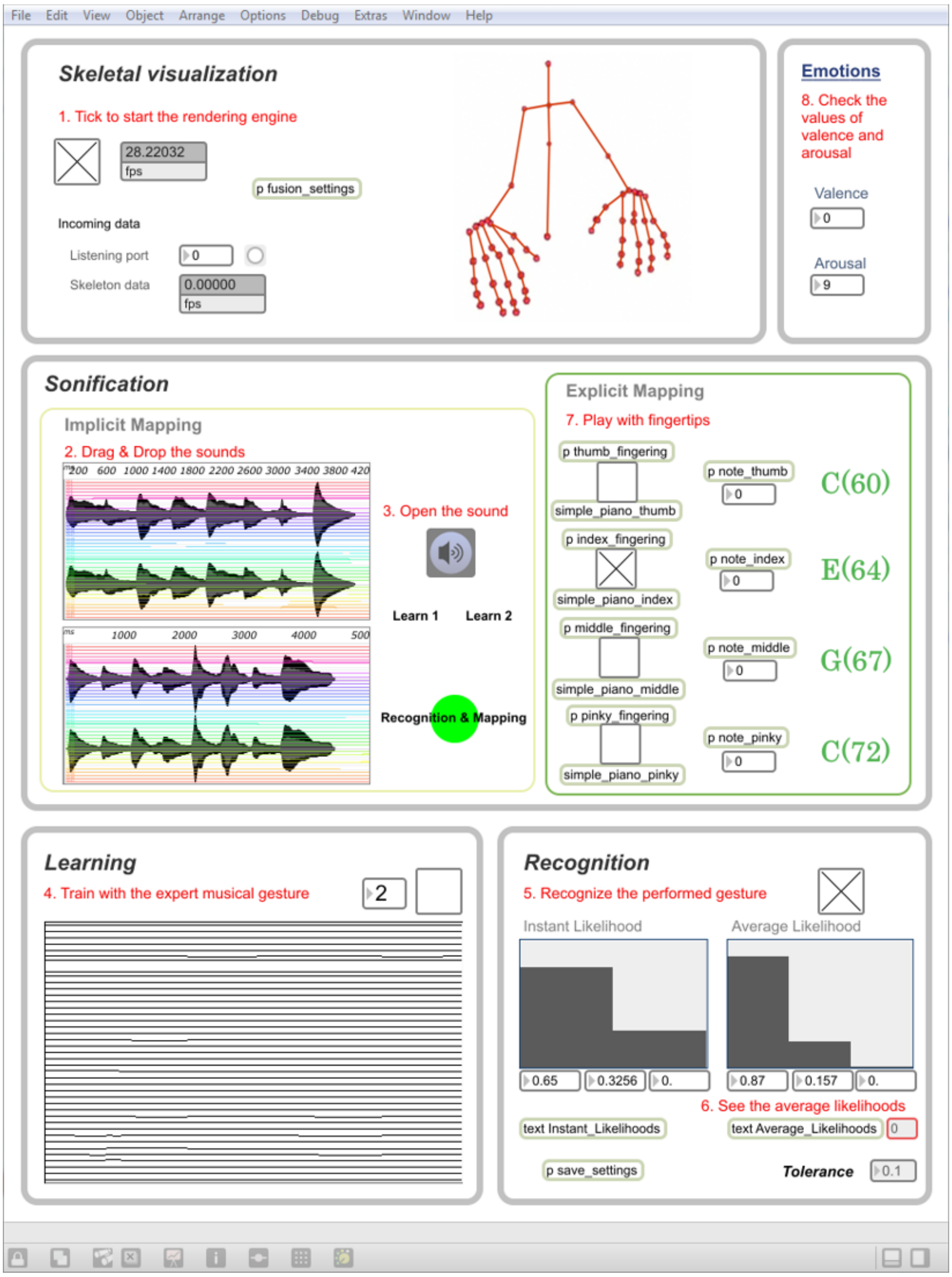


Fig. 2. Intangible Musical Instrument (IMI)

The successive steps, which were achieved in order to first capture the gesture and then to model it, are described below. For the capturing part, two types of depth camera and two inertial sensors are used. The first type of depth camera is the Kinect[3], originally created for video gaming purposes. Equipped with a structured light projector, it can track the movement of the whole body of individuals in 3D using a Random Decision Forest algorithm [Shotton et al. 2013]. However, the proposed methodology focuses on the upper part of the body and the current algorithm delivers a fairly accurate tracking of the head, shoulders, elbows and the hands, but not the fingers. The second type of camera used is the Leap Motion[4], which works with two monochromatic cameras and three infrared LEDs. The Leap Motion provides an accurate description of the hand skeleton, with more than 20 joints positions and velocities, both in 3D (x, y, z coordinates). Two Leap Motion are used, one for each hand. Each Leap Motion has a field of view of 150° and tracks the hand from below efficiently up to 30 cm above the camera center (the camera is oriented upwards). Once placed on their slots on the IMI, they cover the whole surface of the table and a volume above it. Additionally, two inertial sensors are taped to the user's wrists (Animazoo motion capture suit[5]), which deliver rotation angles (Euler angles).

[2] https://www.cycling74.com/
[3] https://www.microsoft.com/en-us/kinectforwindows/
[4] https://www.leapmotion.com/
[5] http://synertial.com/

Finally, an electroencephalogram is mounted on the head to record brain electrical patterns via the Emotiv sensor[6]. These patterns are then translated to the emotional status of the user.



1. Tick to start the rendering engine

28.22032 fps

Listening port 0

Skeleton data 0.00000 fps

8. Check the values of valence and arousal

Valence 0

Arousal 9

2. Drag & Drop the sounds

3. Open the sound

7. Play with fingertips

4. Train with the expert musical gesture 2

5. Recognize the performed gesture

Instant Likelihood: 0.65 0.3256 0.

Average Likelihood: 0.87 0.157 0.

6. See the average likelihoods

Tolerance 0.1

Fig. 3. Unified interface for gesture and emotion recognition and sonification

[6] https://emotiv.com/

In a live learning scenario of the proposed methodology, which is a "learning by doing" approach, the learner has to stand in front of the capture system (Kinect and Leap Motions) and wear the two inertial sensors on his/her wrists and the Emotiv sensor on his/her head, so as to see his/her skeletal representation on the IMI's interface (Skeletal visualization in Figure 3) and attempt to perform the musical expert gestures as well as to embrace the respective emotional status. The expert vocabulary contains some basic musical gestures such as ascending and descending scales, ascending and descending arpeggios, as well as some basic musical excerpts of Beethoven.

In order to perform i.e. ascending and descending scales, the learner's gestural data (positions and Euler angles) are analyzed and used for the machine-learning phase, which is based on HMMs and DTW technique [Bevilacqua et al. 2007; Bevilacqua et al. 2010]. The main advantage of this hybrid approach, instead of using other algorithms, is that it permits a time alignment between the model and the data used as input for the recognition. The two phases are Training (or Learning) and Recognition. In the Training Phase, the expert trains the system with his/her musical gesture, and a pre-recorded sound is associated with the template gesture and links the sound with temporal states of the template gesture (Learning in Figure 3). In the Recognition Phase, the learner tries to imitate in real-time the expert's musical gesture. The meaning of real-time performance and recognition is that the technique does not recognize the gesture once it is completed, but it estimates the gesture in real-time, moment by moment over time. As a result, it is designed to continuously output information about the gesture, by providing the learner's probabilistic estimations (Recognition in Figure 3). Simultaneously, the sonification is taking place based on granular sound synthesis engine, in which the system predicts the sound according to the performed gesture (Implicit Mapping in Figure 3). Moreover, the sound can be modified according to the values of valence and arousal, meaning the emotional status of the user (Emotions in Figure 3). The values of arousal modify the loudness and the values of valence the pitch of the sound. Finally, the learner has the ability to play a musical sequence (i.e. ascending and descending arpeggios), in which each fingertip is associated to each specific note (Explicit Mapping in Figure 3).

5. EVALUATION

This section deals with the evaluation of the IMI and its functionalities in terms of complying with the user's requirements, expectations and experiences. The survey instrument is a structured questionnaire that has been distributed during three scheduled demos in Greece, in which 105 respondents took part. The demographics of the respondents are presented in Table II.

Table II. Demographics of respondents (n=105)

	Number	Percentage (%)
Sex:		
Male	31	29.5
Female	74	70.5
Age:		
Up to 20	21	20.0
21 – 30	36	34.3
31 – 40	20	19.0
41 - 50	28	26.7
Perceived familiarity in using computers:		
Not much	8	7.6
Adequate	66	62.9
Good	31	29.5
Music Literacy:		
Not at all	50	47.6
Not much	39	37.1
Adequate	12	11.4

Good	4	3.8
Familiarity with classical music:		
Not at all	48	45.7
Not much	28	26.7
Adequate	21	20.0
Good	8	7.6

A common rule for considering whether a sample size is acceptable is the ratio of sample size to the number of the latent variables parameters to be equal to 5 to 1 [Bentler and Chou 1987]. Taking into consideration that a sample size that follows this rule is equal to at least 90, it is concluded that the sample size of this study is acceptable. Figure 4 shows the workshops in progress, in which users experiment with IMI while learning and performing musical gestures.

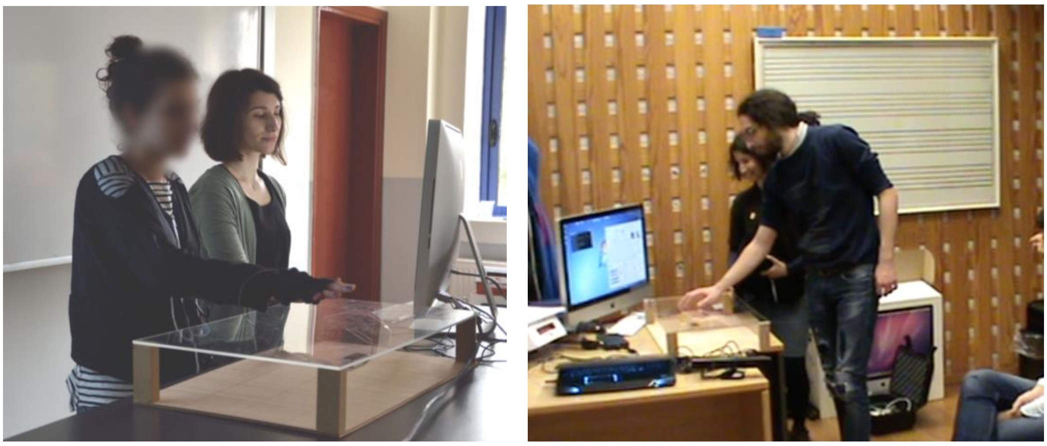


Fig. 4. IMI workshops where the users experiment with the IMI in both learning and performing/composing contexts

During the workshops, the researchers presented the IMI to the participants and a number of expert musicians performed the musical expert gestures that are described in Section 4. Each of those experts used his/her personal musical style to perform on the IMI. Then, participants (users) were asked to identify basic referential elements, meaning musical gestures of the experts and each participant tried to imitate the gestures of his/her favorite expert on the IMI in order to control the expert sound. Figure 5 presents the operational model that was used. Operational model is an abstract and visual representation of how an activity is working, or in other words, it is the blueprint of this activity. Each “box” refers to a general construct, which is constituted by items (questions). Constructs were proposed by expert musicians, teachers and engineers, thus verifying content validity. Special attention was given to basic users, without any specific knowledge of music, in order to verify whether the IMI facilitates the learning of the expert gestures.

The key concept of this operational model argues that the skills of the user on recognizing referential stylistic elements or even specific movement patterns of a given expert musician, mediate the relationship between the quality of interaction with gesture sonification and the performance of the user when playing on the IMI. Based on this operational model the specific goals of this evaluation may be summarized as follows:

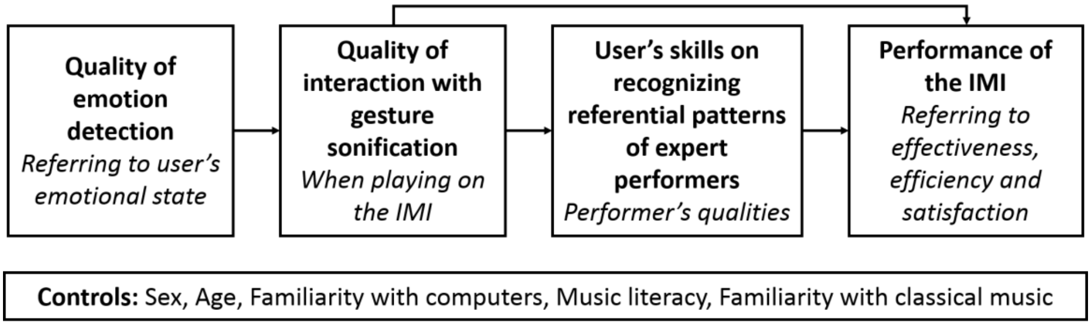

Quality of emotion detection
Referring to user's emotional state
Quality of interaction with gesture sonification
When playing on the IMI
User's skills on recognizing referential patterns of expert performers
Performer's qualities
Performance of the IMI
Referring to effectiveness, efficiency and satisfaction
Controls: Sex, Age, Familiarity with computers, Music literacy, Familiarity with classical music

Fig. 5. The IMI operational model indicating the relationship between quality of emotion detection and IMI performance

G1. To evaluate the overall perceived performance of IMI, with respect to effectiveness (i.e. if IMI meets its objectives), efficiency (i.e. if IMI responses satisfactorily and in a short time in gestures, emotions and sound production), and satisfaction (i.e. if IMI provides satisfaction to the user).

G2. To evaluate whether the personal style (in terms of gesture, music style, and emotions) of a performer (while s/he interprets classical or contemporary composers) can be recognized.

G3. To evaluate the usability and the user-friendliness of the IMI in terms of "outer interactions", such as hands, playing, and setup, and "inner interactions", such as freedom, expression, feedback, motivation, and learning (see Section 5.1).

G4. To evaluate whether the user's experience on emotion detection, such as images, timing, and colors, influences the overall perceived performance of IMI, through the quality of interaction with gesture sonification and the users' perception in recognizing the performer's personal style.

G5. To estimate the total influence of each entity on the overall perceived performance of IMI.

5.1 Validating the IMI operational model

Firstly, exploratory factor analysis (EFA) was performed in order to investigate the dimensions of the constructs proposed. All constructs were uni-dimensional, except for the construct with respect to the interaction with gesture sonification, which produced two dimensions (factors). These dimensions have been labelled "outer interactions", including items such as placing hands (loading = 0.901), playing comfortability (0.847), and setup environment to perform (0.624), and "inner interactions", including items such as freedom (0.656), expression (0.764), audio feedback (0.578), visual feedback (0.690), motivation (0.618), and learning (0.718). Furthermore, the Kaiser-Meyer-Olkin (KMO), measuring the sampling adequacy, and the Bartlett's test of sphericity, measuring the appropriateness of factor analysis, was used [Field 2005]. The KMO value found to be equal to 0.776 (i.e. above the critical value of 0.50) and Bartlett's test exact significance equal to 0.000 (i.e. bellow the critical value of 0.05). These findings, taking also into consideration the corresponding scree plot presented in Figure 6, indicated that factor analysis is appropriate for these data [Kaiser 1974]. In addition, the values of the estimated Cronbach Alphas (above 0.70) and the percentage of the total variance (above 50%), explained from factor analysis for each construct, verified the consistency of the survey instrument and the instrument content validity.

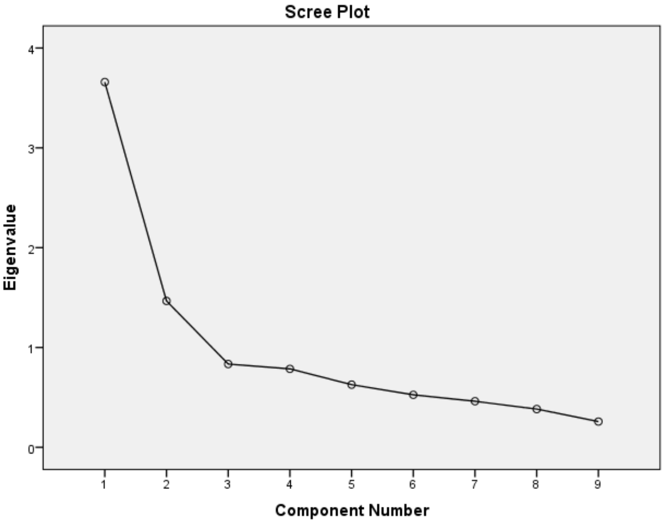


Fig. 6. Scree plot of factor analysis

Furthermore, a joint confirmatory factor analysis (CFA) of all constructs as well as the Kolmogorov – Smirnov normality test [Smirnov 1948] for each construct individually were performed and then the operational model was estimated. The CFA test indicated that the constructs could be used in the estimation of the operational model, in the form presented in Figure 5. The normality tests indicated that all the constructs followed normal distribution patterns and thus the maximum likelihood test could be used in the estimation of the operational model.

Table III presents the means and the standard deviations of all the constructs used in the study, and displays their bivariate correlation coefficients. Strong, positive and significant correlations between the variables involved are observed, supporting the hypotheses of the study. However, results based on correlations, although interesting, may be misleading due to the interactions between several variables [Katou et al. 2014].

Table III. Means, standard deviations and bivariate correlation coefficients of all the constructs

Constructs	Mean (standard deviation)	Correlation Coefficients				
		Emotion detection	Outer Interaction	Inner Interaction	Skills on recognizing	Performance of IMI
Emotion detection	3.025 (0.753)	1				
Outer Interaction	3.483 (0.795)	0.366**	1			
Inner Interaction	2.981 (0.711)	0.475**	0.439**	1		
Skills on Recognizing	3.127 (0.686)	0.421**	0.431**	0.561**	1	
Performance of IMI	3.406 (0.794)	0.465**	0.382**	0.539**	0.455**	1

**. Correlation is significant at the 0.01 level (2-tailed).

Therefore, in order to isolate the possible links between the variables involved in the operational model presented in Figure 5, the estimated path diagrams for this proposed framework are presented in Figure 7. The boxes represent exogenous or endogenous observed variables and the circles represent the related latent variables. The light arrows indicate the observed variables that constitute the related latent variables and the bold arrows indicate the structural relationships between the corresponding variables. For comparison purposes, the numbers that are assigned to each arrow show the estimated standardized coefficients. However, under the structural estimated standardized coefficients, the numbers in brackets present the actual estimated coefficients and their standard errors, indicating that the confidence intervals for the estimated coefficients are very narrow.

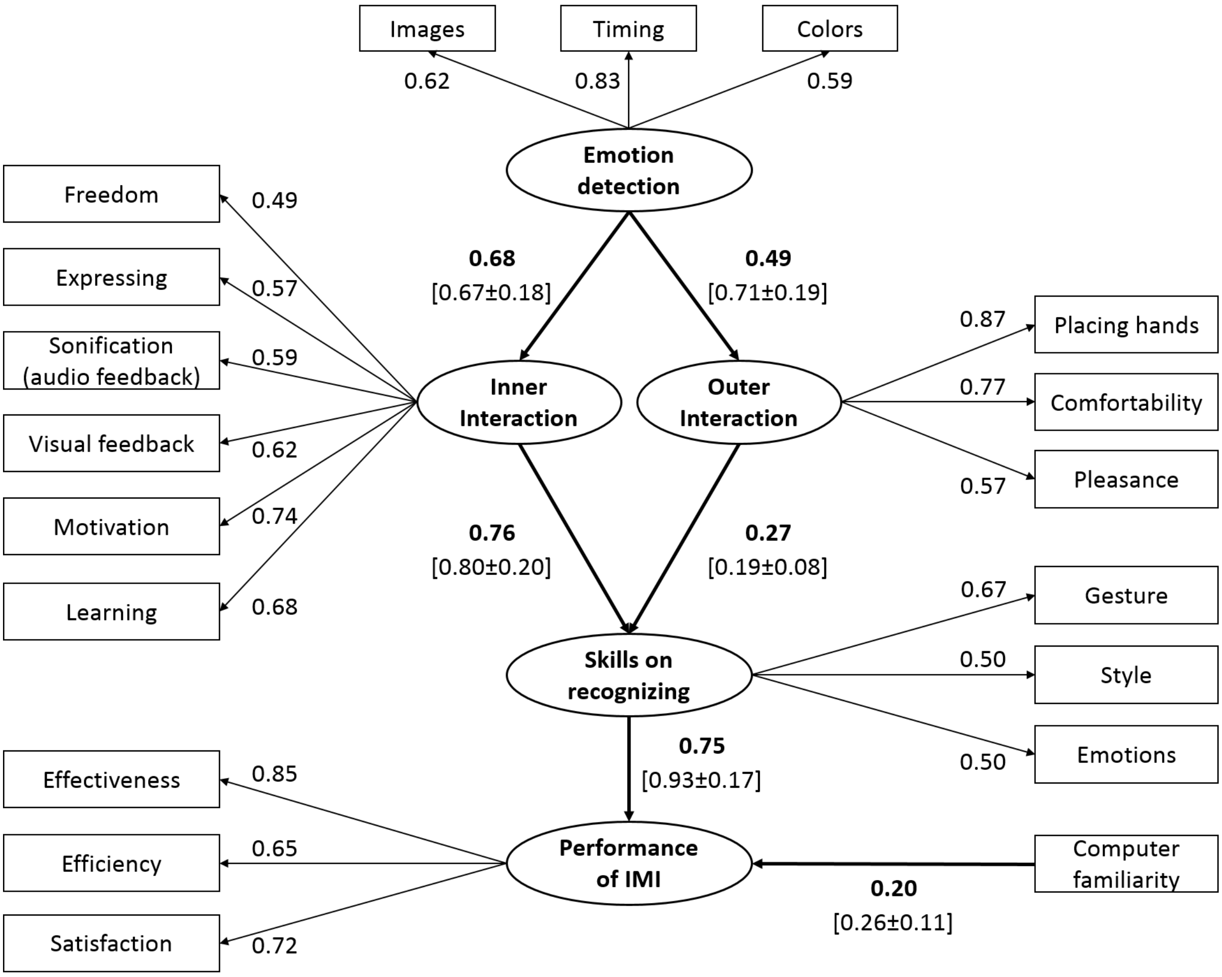


Fig. 7. Estimation results of the IMI operational model

Table IV presents the fit indices that are attached to the results presented in Figure 7. The performance of the IMI operational model is very satisfactory as it can predict the IMI processes with 93.7% overall accuracy. Taking into account that chi-square statistics may be inflated by significant correlations between constructs, the value of the normed-chi-square was used instead. In our case this value is very small (1.412), confirming the validity of our model and indicating that the proposed model is an adequate presentation of the entire set of relationships [Pedhazur and Pedhazur-Schelkin 1991]. In addition, the values of the GFI and the CFI are above the corresponding critical values, verifying that the structure of the model fits the empirical data satisfactorily. However, the value of the NFI (0.738) is much less than its critical value, indicating the usual tendency to underestimate fit in relatively small samples [Bentler 1990].

Table IV. Assessment indices of the IMI operational model

Assessment category	Fit Indices			
	ID	Description	Pass criteria	Value
Overall Performance Evaluation Indices	Chi-square	Exact significance of Chi-square statistic	>0.05	0.001
	Normed Chi-square	Chi-square / Degrees of freedom	<5	1.412
	RMSEA	Root Mean Squared Error Approximation (numerical value [0,1])	<0.10	0.063
Individual Fit Indices	GFI	Goodness of Fit Index (numerical value [0,1])	>0.70	0.837
	NFI	Normed Fit Index (numerical value [0,1])	>0.90	0.738
	CFI	Comparative Fit Index (numerical value [0,1])	>0.90	0.903

5.2 Evaluation results

The above results highlight that the model satisfactorily predicts performance (all standardized coefficients are significant and positive, and fit indices are acceptable overall). Moreover, the skills on recognizing gestures, styles and the emotions of a performer fully mediate [Baron and Kenny 1986] the relationship between interaction and performance. The ease for placing of hands, degree of comfort, motivation, and learning are the most important interactional factors with IMI. In terms of emotional elicitation, the results highlighted the timing of the user's emotional state as the most important factor. Furthermore, gesture recognition is the most important factor in determining performance effectiveness. Performance is also influenced by familiarity in using computers (all the other controls, such as sex, age, educational level, music literacy, used in the study were not significant). This means that IMI belongs to the so-called "Contingency Systems", which support the view that the system maximizes its performance according to the specific context in which it is operating [Delery and Doty 1996]. However, it should be taken into consideration that the model and its estimation is based on perceived subjective data. Perceived data do not undermine the usefulness of the model and the whole IMI evaluation exercise. In any case, a further technical assessment exercise could also be employed to measure performance of the system in a more objective manner.

Overall, Table V presents the total effect of each column variable on each row variable after standardizing all variables. For example, the standardized total (direct and indirect) effect of the "emotion detection" on satisfaction is 0.347. That is, due to both direct (unmediated) and indirect (mediated) effects of the "emotion detection" on satisfaction, when it goes up by 1 standard deviation, satisfaction goes up by 0.347 standard deviations (for further discussion of direct, indirect and total effects, see Kline [Kline 1998]).

Table V. Total standardized effects for IMI operational model

ID of Question (refers to the items in the Questionnaire in Appendix)	Variable Abbreviation	Constructs				
		Emotion detection	Inner Interaction	Outer Interaction	Skills on recognizing	Performance of IMI
	Inner Interaction	0.675	0.000	0.000	0.000	0.000
	Outer Interaction	0.489	0.000	0.000	0.000	0.000
	Skills on recognizing	0.644	0.761	0.267	0.000	0.000
	Performance of IMI	0.482	0.569	0.200	0.748	0.000
Q1.1.1	Hands	0.423	0.000	0.865	0.000	0.000

Q1.1.2	Comfort	0.374	0.000	0.766	0.000	0.000
Q1.1.3	Pleasant	0.277	0.000	0.566	0.000	0.000
Q1.1.4	Freedom	0.333	0.493	0.000	0.000	0.000
Q1.1.5	Expression	0.387	0.573	0.000	0.000	0.000
Q1.1.6	Sonification	0.396	0.586	0.000	0.000	0.000
Q1.1.7	Visual	0.418	0.619	0.000	0.000	0.000
Q1.1.8	Motivation	0.500	0.740	0.000	0.000	0.000
Q1.1.9	Learning	0.459	0.680	0.000	0.000	0.000
Q1.2.1	Colors	0.588	0.000	0.000	0.000	0.000
Q1.2.2	Time	0.835	0.000	0.000	0.000	0.000
Q1.2.3	Images	0.616	0.000	0.000	0.000	0.000
Q2.1	Gesture	0.429	0.506	0.178	0.666	0.000
Q2.2	Style	0.319	0.377	0.132	0.496	0.000
Q2.3	Emotions	0.322	0.381	0.134	0.500	0.000
Q3.1	Effectiveness	0.408	0.482	0.169	0.634	0.848
Q3.2	Efficiency	0.313	0.369	0.130	0.485	0.649
Q3.3	Satisfaction	0.347	0.410	0.144	0.539	0.721

Table V could guide the designers of the IMI to put effort into improving entities in the system in order to get better results, as perceived by the users of the system; however, any amendment should take into consideration the cost-benefit analysis of each change.

A summary of Table V is presented in Figure 8, where the relationship between the components (i.e. explanatory variables) and performance (i.e. dependent variable) of IMI is presented. This summary indicates both the mean values of each component and their contribution (loadings) to the performance of IMI. Thus, any amendment/update of the components will influence performance.

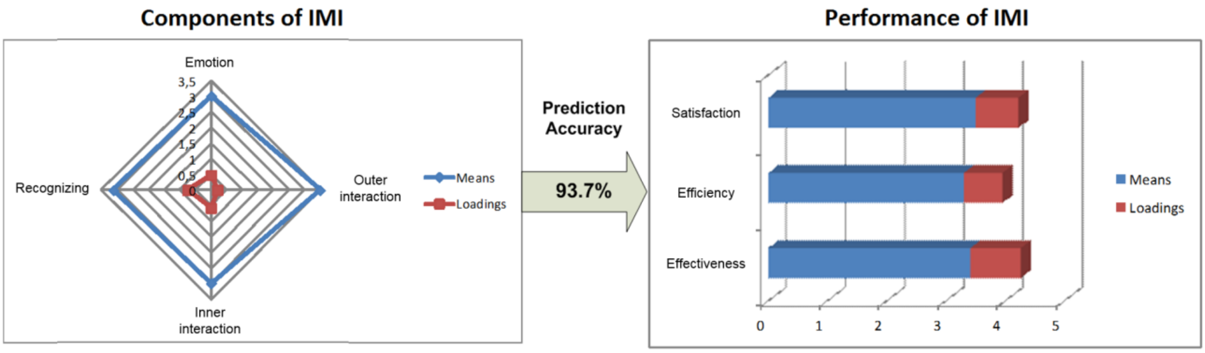


Fig. 8. The relationship between components of IMI and its performance

For amending/updating the components of IMI, the following rules should be considered:

(1) the prediction accuracy index, explaining the relationship between the explanatory variables (components of IMI) and the dependent variable (dimensions of performance) should be as close as possible to one (perfect prediction);
(2) the means of the variables (explanatory and/or dependent) should be as close to level five (perfect perceived user's evaluation response); and
(3) the loadings of the variables (explanatory and/or dependent) should be as close to one (perfect contribution to performance).

Additionally, the following information are also used:

a) a low mean of a variable means that there is "room" for the corresponding component to be improved; however, a cost-benefit analysis should also be used to investigate how easy it is to improve a component;

b) the high loading of a variable means that the corresponding component is important in determining performance.

This means that in ranking the amendments/updates of the components of IMI, it should be taken into consideration whether there is "room for improvement" and whether there is a "high contribution". This can be achieved by considering the following general rule, which accordingly determines a "component ranking index": "The lower the ratio of the mean of a component is, the higher the priority for amendments/updates for this component".

Following this rule, the major conclusions and recommendations with respect to IMI can now be summarized. First of all, the components of IMI predict performance with very high accuracy (93.7%). The perception of the dimensions of IMI performance range between 3.30 and 3.50 (on a five-level Likert scale), indicating that the performance of IMI is above average. According to the component ranking index, the components of the IMI, in a descending order, are arranged as follows: outer interaction with gesture sonification (17.40), emotion detection (6.27), inner interaction with gesture sonification (5.24), and skills on recognizing performer's qualities (4.17). Considering the arrangement in (3), it is recommended that the priority for improvements to be taken should be in the areas of "skills on recognizing performer's qualities" and "inner interaction with gesture sonification". Applying the same rule within the components of "skills on recognizing performer's qualities", priority for improvements should be taken in the areas of gestures (4.58), style (6.26), and emotions (6.36). Applying further the same rule within the components of "inner interaction with gesture sonification", the priority for improvements should be taken in the areas of motivation (3.96), visual feedback (4.68), audio feedback/sonification (4.95), expressing (4.84), learning (5.00), and freedom (6.04).

6. CONCLUSIONS

A prototype natural user interface, named the Intangible Musical Instrument (IMI), was presented in this paper. From a technical point of view, this first prototype is able to capture, model and recognize musical gestures and emotions as well as to sonify gestures and emotions. The IMI is conceived to transmit the multi-layer musical ICH to the public, by developing a unified interface framework that supports learning, performing and composing with gestures. This means that the learning of gestural knowledge of expert performers becomes a first-person experience with the IMI. Most importantly, a significant effort has been made to put the user at the core of the musical performance and of composition with gestures in both combined and autonomous ways.

Summarizing, the first evaluation of the IMI shows very satisfactory results in performance prediction. However, what emerges from the study described here is that future work should focus more on improvements in terms of research into expert musical gesture recognition as well as visual and audio feedback (sonification). Our future research should also focus on augmenting ordinary musical scores by providing gestural and emotional annotations together with the musical notation, in order to further facilitate the learning experience.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union, Seventh Framework Programme (FP7-ICT-2011-9) under grant agreement n° 600676. We would also like to thank Frédéric Bevilacqua for giving us his permission to use the 'Gesture Follower' in testing our hypothesis.

REFERENCES

Reuben M. Baron and David A. Kenny. 1986. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–82.

Peter M. Bentler. 1990. Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.

Peter M. Bentler and Chih-Ping Chou. 1987. Practical issues in structural modeling. *Sociological Methods & Research*, 16(1), 78-117.

Frédéric Bevilacqua, Fabrice Guédy, Nobert Schnell, Emmanuel Fléty and Nicolas Leroy. 2007. Wireless sensor interface and gesture-follower for music pedagogy. In *Proceedings of the International Conference of New Interfaces for Musical Expression*, New York, USA, 124-129.

Frédéric Bevilacqua, Bruno Zamborlin, Anthony Sypniewski, Norbert Schnell, Fabrice Guédy and Nicolas Rasamimanana. 2010. Continuous realtime gesture following and recognition, *LNAI* 5934, 73–84.

Frédéric Bevilacqua, Norbert Schnell, Nicolas Rasamimanana, Bruno Zamborlin and Fabrice Guédy. 2011. *Online gesture analysis and control of audio processing*. In: J. Solis & K.C. Ng (Eds.). Musical Robots and Interactive Multimodal Systems (Springer Tracts in Advanced Robotics, 74, 127–142). Berlin, Heidelberg: Springer.

Jan L. Broeckx. 1981. Muziek, ratio en affect over de wisselwerking van rationeel denken en affectief beleven bij voortbrengst en ontvangst van muziek. Antwerpen: Metropolis.

Bernd Buxbaum. 2002. Optische Laufzeitmessung und CDMA auf Basis der PMD-Technologie mittels phasenvariabler PN-Modulation, Schaker Verlag, Aachen.

Claude Cadoz and Marcelo M. Wanderley. 2000. *Gesture-music. Trends in gestural control of music*. In: M.M. Wanderley and M. Battier (Eds.), Trends in gestural control of music. Paris, IRCAM/Centre Pompidou, 71–94.

Antonio Camurri, Gualtiero Volpe, Giovanni de Poli and Marc Leman. 2005. Communicating Expressiveness and Affect in Multimodal Interactive Systems. *IEEE MultiMedia*, 12(1), 43-53.

Martin R. L. Clayton. 2000. Time in Indian Music: Rhythm Metre and Form in Indian Rag Performance. Oxford: Oxford University Press.

Wilson Coker. 1972. Music & Meaning: A Theoretical Introduction to Musical Aesthetics. New York: The Free Press.

Naomi Cumming. 2000. *The sonic self: Musical subjectivity and signification*. Bloomington: Indiana University Press.

Francois Delalande. 1988. *La gestique de Gould: éléments pour une sémiologie du geste musical*. In: G. Guertin (Eds.) Glenn Gould pluriel, Louise Courteau éditrice, Montréal.

John Delery and Harold D. Doty. 1996. Modes of theorizing in strategic human resource management: test of universalistic, contingency and configurational performance predictions. *Academy of Management Journal*, 39, 802-835.

Andy P. Field. 2005. *Discovering statistics using SPSS* (2nd edition). London: Sage.

Anders Friberg and Johan Sundberg. 1999. Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America*, 105(3), 146-148.

Aalf Gabrielsson and Erik Lindström. 2010. *The role of structure in the musical expression of emotions*. In: P.N. Juslin & J.A. Sloboda (Eds.). Handbook of music and emotions theory, research, applications. New York, NY: Oxford University Press, 367-400.

Leontios J. Hadjileontiadis. 2014. Conceptual Blending in Biomusic Composition Space: The "Brainswarm" Paradigm. In *Proceedings of the ICMC/ SMC Conference*. Athens.

Robert S. Hatten. 1994. Musical meaning in Beethoven: markedness, correlation, and interpretation. Bloomington: Indiana University Press.

Rolf Inge Godøy and Mark Leman. 2009. *Musical gestures: Sound, Movement, and Meaning*. In: Rolf Inge Godoy and Marc Leman (Eds.), Routledge, New York.

Alexander R. Jensenius. 2007. *ACTION – SOUND Developing Methods and Tools to Study*. Ph.D. Dissertation, University of Oslo.

Patrik N. Juslin. 2003. Five facets of musical expression: A psychologist's perspective on music performance, *Psychology of Music*, 31(1), pp. 273-302.

Patrik N. Juslin and John Sloboda. 2010. *Handbook of music and emotions theory, research, applications*. New York, NY: Oxford University Press.

Patrik N. Juslin and Renee Timmers. 2010. *Expression and communication of emotion in music performance*. In: P.N. Juslin & J.A. Sloboda (Eds.). Handbook of music and emotions theory, research, applications. New York, NY: Oxford University Press, 453-489.

Henry F. Kaiser. 1974. An index of factorial simplicity. *Psychometrika*, 39, 31-36.

Anastasia A. Katou, Pawan S. Budhwar and Charmi Patel. 2014. Content vs. process in the HRM-performance relationship: An empirical examination. *Human Resource Management*, 53(4), 527-544.

Peter Keller. 2008. *Joint action in music performance*. In: F. Morganti, A. Carassa and G. Riva (Eds.). Enacting Intersubjectivity: A Cognitive and Social Perspective on the Study of Interaction. Amsterdam, 205-221.

Rex B. Kline. 1998. Principles and practice of structural equation modeling. NY: Guilford Press.

Stefan Koelsch and Walter A. Siebel. 2005. Towards a neural basis of music perception, *Trends in Cognitive Sciences*, 9(12), 578 – 584.

Marc Leman. 2010. Music, Gesture, and the Formation of Embodied Meaning. *Musical gestures: Sound, Movement, and Meaning*. 126-153.

Pieter-Jan Maes, Marc Leman, Micheline Lesaffre, Michiel Demey and Dirk Moelants. 2010. From expressive gesture to sound.

The development of an embodied mapping trajectory inside a musical interface. *Journal on Multimodal User Interfaces*, 3(1), 67-78.

David Mcneill. 1992. *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL: University of Chicago Press.

Alva Noë. 2004. *Action in Perception*. Cambridge, MA: MIT Press.

Elazar J. Pedhazur and Liora Pedhazur-Schmelkin. 1991. Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum.

Carroll C. Pratt. 1931/1968. The meaning of music: A study in psychological aesthetics, New York: Johnson.

James A. Russell. 1980. A circumflex model of affect, *Journal of Personality and Social Psychology*, 39, 1161-1178.

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman and Andrew Blake. 2013. Real-Time Human Pose Recognition in Parts from a Single Depth Image. *Machine Learning for Computer Vision*, SCI, 411, 119-135.

Alexander Truslit. 1938. *Gestaltung und Bewegung in der Musik*. Berlin-Lichterfelde: C.F. Vieweg.

N. Smirnov. 1948. Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics, 19, 279–281.

UNESCO 2003. Convention of the Safeguarding of Intangible Cultural Heritage of UNESCO. Available: https://ich.unesco.org/en/convention.

Frank Weichert, Daniel Bachmann, Bartholomäus Rudak and Denis Fisseler. 2013. Analysis of the accuracy and robustness of the leap motion controller. *Sensors*, 13, 6380–6393.

Matthias Wiedemann, Markus Sauer, Frauke Driewer and Klaus Schilling. 2008. Analysis and characterization of the PMD camera for application in mobile robotics. In *Proceedings of the 17th IFAC World Congress*, 6-11.

Liwei Zhao. 2001. Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures. Ph.D. Dissertation. University of Pennsylvania, Philadelphia, PA, USA. AAI3015399.

Appendix:

SECTION 1: QUALITY CHARACTERISTICS OF INTANGIBLE MUSICAL INSTRUMENT

1. How would you rate the interaction with IMI?

No.	**Interactive Characteristics**	Not at all				Very much
		1	2	3	4	5
1	How easy did you find it to place your hands correctly (correct octave) on the IMI?					
2	How comfortable did you find the playing/performance of the musical gestures on the IMI?					
3	How pleasant did you find the setup environment to perform with in terms of design and aesthetics (Plexiglas, motion sensors, brain activity sensors)?					
4	Did you feel that your gestures were more free compared to a real piano?					
5	To what extent could you express yourself through the IMI?					
6	How much did the IMI help you to improve/correct your gesture during performance, by comparing the sound that you produced in real-time (sonification/audio feedback) with the sound that you listened to while watching the video with the expert's gestures?					
7	Did the virtual avatar (visual feedback) help you improve/correct the performance of your gestures?					
8	How much did the IMI activity of "Final Challenge" motivate you to focus more on the learning of fundamental musical gestures and activities?					
9	How much do you think the IMI would help you to learn musical gestures?					

2. How would you rate the quality of emotion detection?

No.	**Emotions**	Very Bad				Very good
		1	2	3	4	5
1	To what extent did the images that were shown to you excite you emotionally?					
2	To what extent was the time enough for you to follow the emotional state at every level of the game?					
3	To what extent were the colors representative of emotions?					

SECTION 2: USER'S MUSIC PERCEPTION

How easily can you recognize the personal style (expressiveness, gestures, sound) of a performer while s/he interprets classical or contemporary composers?

No.	**Performer's qualities**	Very Bad				Very Good
		1	2	3	4	5
1	Gestures					
2	Music Style (*e.g. to understand differences between playing/performance while listening to the same piece of music*)					
3	Emotions					

SECTION 3: PERFORMANCE OF THE IMI

How would you rate the overall performance of the IMI?

No.	**Performance Measures**	Very Bad				Very Good
		1	2	3	4	5
1	**Effectiveness** (*if the IMI meets its objectives*)					
2	**Efficiency** (*if the IMI responds satisfactorily and in a short time to gestures, emotions and sound production*)					
3	**Satisfaction** (*if the IMI provides satisfaction to the user*)					

Thank you very much for your co-operation.

List of Acronyms

AI	Artificial Intelligence	i
UAV	Unmanned Aerial Vehicle	6
AGV	Automated Guided Vehicle	11
HAI	Human-centred Artificial Intelligence	2
IMU	Inertial Measurement Unit	12
CNN	Convolutional Neural Network	14
DTW	Dynamic Time Warping	13
GOM	Gesture Operational Model	10
SS	State-Space	13
MLE	Maximum Likelihood Estimation	10
HMM	Hidden Markov Model	11
NIME	New Interfaces for Musical Expression	80
HRC	Human-Robot Collaboration	7
PCA	Principal Component Analysis	73
DMI	Digital Musical Instrument	i
PASR	preparation, attack, sustain, release	89

GVF Gesture Variation Follower 13

GF Gesture Follower 12

BPM Beats Per Minute 94

FFT Fast Fourier Transformation 99

CCI Creative and Cultural Industries 4

GMM Gaussian Mixture Model 63

RNN Recurrent Neural Network 14

GAIL Generative Adversarial Imitation Learning 64

LCSS Longest Common Subsequence 65

DDTW Derivative Dynamic Time Warping 65

ML Machine Learning 42

DL Deep Learning 11

WDTW Weighted Dynamic Time Warping 65

pHRC physical Human-Robot Collaboration 41

tHRC touchless Human-Robot Collaboration 41

WMSD Work-related Musculoskeletal Disorders 105

EAWS European Assessment Worksheet 105

RULA Rapid Upper Limb Assessment 105

SSI Silent Speech Interface 7

TTS Text-to-Speech 104

FoF Factory of the Future 4

IoT Internet of Things 1

KPI Key Performance Indicator 106

LSTM Long Short-Term Memory 16

Bibliography

[Adams 1971] Jack A Adams. *A closed-loop theory of motor learning.* Journal of Motor Behavior, vol. 3, no. 2, pages 111–150, 1971. (Cited on page 65.)

[Akkaladevi *et al.* 2018] Sharath Chandra Akkaladevi, Matthias Plasch, Sriniwas Maddukuri, Christian Eitzinger, Andreas Pichler and Bernhard Rinner. *Toward an Interactive Reinforcement Based Learning Framework for Human Robot Collaborative Assembly Processes.* Frontiers in Robotics and AI, vol. 5, page 126, 2018. (Cited on page 43.)

[Anderson *et al.* 1996] John R Anderson, Lynne M Reder and Herbert A Simon. *Situated learning and education.* Educational Researcher, vol. 25, no. 4, pages 5–11, 1996. (Cited on page 2.)

[Asadi-Aghbolaghi *et al.* 2017] Maryam Asadi-Aghbolaghi, Albert Clapes, Marco Bellantonio, Hugo Jair Escalante, Víctor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei and Sergio Escalera. *A survey on deep learning based approaches for action and gesture recognition in image sequences.* In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pages 476–483. IEEE, 2017. (Cited on page 15.)

[Asaula *et al.* 2010] Ruslan Asaula, Daniele Fontanelli and Luigi Palopoli. *Safety provisions for human/robot interactions using stochastic discrete abstractions.* In 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 2175–2180. IEEE, 2010. (Cited on page 43.)

[Astill & Utley 2008] Sarah Astill and Andrea Utley. *Coupling of the reach and grasp phase during catching in children with developmental coordination disorder.* Journal of Motor Behavior, vol. 40, no. 4, pages 315–324, 2008. (Cited on page 65.)

[Bakis 1976] Raimo Bakis. *Continuous speech recognition via centisecond acoustic states.* The Journal of the Acoustical Society of America, vol. 59, no. S1, pages S97–S97, 1976. (Cited on page 25.)

[Barth & Franke 2008] Alexander Barth and Uwe Franke. *Where will the oncoming vehicle be the next second?* In 2008 IEEE Intelligent Vehicles Symposium, pages 1068–1073. IEEE, 2008. (Cited on page 16.)

[Baum *et al.* 1972] Leonard E Baum*et al.* *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes.* Inequalities, vol. 3, no. 1, pages 1–8, 1972. (Cited on page 25.)

[Best & Fitch 2015] Graeme Best and Robert Fitch. *Bayesian intention inference for trajectory prediction with an unknown goal destination.* In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5817–5823. IEEE, 2015. (Cited on page 17.)

[Bevilacqua *et al.* 2007] Frederic Bevilacqua, Fabrice Guédy, Norbert Schnell, Emmanuel Fléty and Nicolas Leroy. *Wireless sensor interface and gesture-follower for music pedagogy.* In Proceedings of the 7th International Conference on New Interfaces for Musical Expression, pages 124–129, 2007. (Cited on page 68.)

[Bevilacqua *et al.* 2009] Frédéric Bevilacqua, Bruno Zamborlin, Anthony Sypniewski, Norbert Schnell, Fabrice Guédy and Nicolas Rasamimanana. *Continuous realtime gesture following and recognition.* In International Gesture Workshop, pages 73–84. Springer, 2009. (Cited on pages 12, 47 and 68.)

[Binelli *et al.* 2005] Emanuele Binelli, Alberto Broggi, Alessandra Fascioli, Stefano Ghidoni, Paolo Grisleri, Thorsten Graf and M Meinecke. *A modular tracking system for far infrared pedestrian recognition.* In IEEE Proceedings. Intelligent Vehicles Symposium, 2005., pages 759–764. IEEE, 2005. (Cited on pages 15 and 16.)

[Cadoz & Wanderley 2001] Claude Cadoz and Marcelo M Wanderley. *Gesture-music: Trends in gestural control of music.* Paris: IRCAM, 2001. (Cited on page 81.)

[Cai *et al.* 2013] Jun Cai, Thomas Hueber, Sotiris Manitsaris, Pierre Roussel, Lise Crevier-Buchman, Maureen Stone, Claire Pillot-Loiseau, Gérard Chollet, Gérard Dreyfus and Bruce Denby. *Vocal tract imaging system for post-laryngectomy voice replacement.* In 2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), pages 676–680. IEEE, 2013. (Cited on page 104.)

[Calinon *et al.* 2011] Sylvain Calinon, Antonio Pistillo and Darwin G Caldwell. *Encoding the time and space constraints of a task in explicit-duration hidden Markov model.* In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 3413–3418. IEEE, 2011. (Cited on page 12.)

[Calinon 2016] Sylvain Calinon. *A tutorial on task-parameterized movement learning and retrieval.* Intelligent Service Robotics, vol. 9, no. 1, pages 1–29, 2016. (Cited on page 63.)

[Camurri *et al.* 2005] Antonio Camurri, Gualtiero Volpe, Giovanni De Poli and Marc Leman. *Communicating expressiveness and affect in multimodal interactive systems.* IEEE Multimedia, vol. 12, no. 1, pages 43–53, 2005. (Cited on page 82.)

[Canal *et al.* 2018] Gerard Canal, Emmanuel Pignat, Guillem Alenyà, Sylvain Calinon and Carme Torras. *Joining High-Level Symbolic Planning with Low-Level Motion Primitives in Adaptive HRI: Application to Dressing Assistance.* In 2018 IEEE International Conference on Robotics and Automation (ICRA), pages 3273–3278, 2018. (Cited on page 42.)

[Cao *et al.* 2017] Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu and Jian Cheng. *Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules.* In Proceedings of the IEEE International Conference on Computer Vision, pages 3763–3771, 2017. (Cited on page 15.)

[Cao *et al.* 2018] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei and Yaser Sheikh. *OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields.* arXiv preprint arXiv:1812.08008, 2018. (Cited on page 19.)

[Cao *et al.* 2019] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei and Yaser Sheikh. *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*, 2019. (Cited on page 14.)

[Caramiaux *et al.* 2014] Baptiste Caramiaux, Nicola Montecchio, Atau Tanaka and Frédéric Bevilacqua. *Adaptive gesture recognition with variation estimation for interactive systems.* ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 4, no. 4, pages 1–34, 2014. (Cited on pages 63 and 92.)

[Caramiaux *et al.* 2015] Baptiste Caramiaux, Marco Donnarumma and Atau Tanaka. *Understanding gesture expressivity through muscle sensing.* ACM Transactions on Computer-Human Interaction (TOCHI), vol. 21, no. 6, pages 1–26, 2015. (Cited on page 13.)

[Caramiaux *et al.* 2020] Baptiste Caramiaux, Jules Françoise, Wanyu Liu, Téo Sanchez and Frédéric Bevilacqua. *Machine Learning Approaches For Motor Learning: A Short Review.* Frontiers in Computer Science, vol. 2, page 16, 2020. (Cited on pages 62 and 64.)

[Carreira & Zisserman 2017] Joao Carreira and Andrew Zisserman. *Quo vadis, action recognition? a new model and the kinetics dataset.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017. (Cited on page 15.)

[Chadabe 2002] Joel Chadabe. *The limitations of mapping as a structural descriptive in electronic instruments.* In Proceedings of the 2002 Conference on New Interfaces for Musical Expression, pages 1–5. Citeseer, 2002. (Cited on page 80.)

[Chalasani *et al.* 2018] Tejo Chalasani, Jan Ondrej and Aljosa Smolic. *Egocentric gesture recognition for head-mounted AR devices.* In 2018 IEEE International

Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pages 109–114. IEEE, 2018. (Cited on page 14.)

[Chen *et al.* 2020] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu and Siddhartha Srinivasa. *Trust-aware decision making for human-robot collaboration: Model learning and planning.* ACM Transactions on Human-Robot Interaction (THRI), vol. 9, no. 2, pages 1–23, 2020. (Cited on page 43.)

[Cheng *et al.* 2019] Yujiao Cheng, Weiye Zhao, Changliu Liu and Masayoshi Tomizuka. *Human motion prediction using semi-adaptable neural networks.* In 2019 American Control Conference (ACC), pages 4884–4890. IEEE, 2019. (Cited on page 63.)

[Cherubini *et al.* 2016] Andrea Cherubini, Robin Passama, André Crosnier, Antoine Lasnier and Philippe Fraisse. *Collaborative Manufacturing with Physical Human–Robot Interaction.* Frontiers in Neuroscience, vol. 40, pages 1–13, 2016. (Cited on page 42.)

[Chollet *et al.* 2015] François Cholletet *al.* *Keras: Deep learning library for theano and tensorflow.* URL: https://keras. io/k, vol. 7, no. 8, page T1, 2015. (Cited on page 30.)

[Côté-Allard *et al.* 2019] Ulysse Côté-Allard, Cheikh Latyr Fall, Alexandre Drouin, Alexandre Campeau-Lecours, Clément Gosselin, Kyrre Glette, François Laviolette and Benoit Gosselin. *Deep learning for electromyographic hand gesture signal classification using transfer learning.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, no. 4, pages 760–771, 2019. (Cited on page 63.)

[Coupeté *et al.* 2015] Eva Coupeté, Fabien Moutarde and Sotiris Manitsaris. *Gesture recognition using a depth camera for human robot collaboration on assembly line.* Procedia Manufacturing, vol. 3, pages 518–525, 2015. (Cited on page 48.)

[Coupeté *et al.* 2016] Eva Coupeté, Fabien Moutarde, Sotiris Manitsaris and Olivier Hugues. *Recognition of Technical Gestures for Human-Robot Collaboration in Factories.* In The Ninth International Conference on Advances in Computer-Human Interactions, 2016. (Cited on page 48.)

[Coupeté *et al.* 2019] Eva Coupeté, Fabien Moutarde and Sotiris Manitsaris. *Multi-users online recognition of technical gestures for natural human–robot collaboration in manufacturing.* Autonomous Robots, vol. 43, no. 6, pages 1309–1325, 2019. (Cited on page 46.)

[Coupeté 2016] Eva Coupeté. *Reconnaissance de gestes et actions pour la collaboration homme-robot sur chaîne de montage.* PhD thesis, 2016. (Cited on page 38.)

[d'Alessandro *et al.* 2015] Nicolas d'Alessandro, Joëlle Tilmanne, Ambroise Moreau and Antonin Puleo. *Airpiano: A multi-touch keyboard with hovering control.* In Proceedings of the International Conference on New Interfaces for Musical Expression, pages 255–258, 2015. (Cited on page 83.)

[Dapogny *et al.* 2013] Arnaud Dapogny, Raoul De Charette, Sotiris Manitsaris, Fabien Moutarde and Alina Glushkova. *Towards a hand skeletal model for depth images applied to capture music-like finger gestures.* In 10th International Symposium on Computer Music Multidisciplinary Research (CMMR'2013), 2013. (Cited on page 104.)

[Darwin 1872] Charles Darwin. *The expression of emotions in animals and man.* London: Murray, vol. 11, 1872. (Cited on page 82.)

[de Charette & Manitsaris 2019] Raoul de Charette and Sotiris Manitsaris. *3d reconstruction of deformable revolving object under heavy hand interaction.* arXiv preprint arXiv:1908.01523, 2019. (Cited on page 104.)

[Delalande 1988] François Delalande. *La gestique de Gould: éléments pour une sémiologie du geste musical.* Glenn Gould Pluriel, pages 85–111, 1988. (Cited on page 81.)

[Dempster *et al.* 1977] Arthur P Dempster, Nan M Laird and Donald B Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, no. 1, pages 1–22, 1977. (Cited on page 25.)

[Denby *et al.* 2011] Bruce Denby, Jun Cai, Thomas Hueber, Pierre Roussel, Gérard Dreyfus, Lise Crevier-Buchman, Claire Pillot-Loiseau, Gérard Chollet, Sotiris Manitsaris and Maureen Stone. *Towards a practical silent speech interface based on vocal tract imaging.* In 9th International Seminar on Speech Production (ISSP 2011), pages 89–94, 2011. (Cited on page 104.)

[Devineau *et al.* 2018] Guillaume Devineau, Fabien Moutarde, Wang Xi and Jie Yang. *Deep learning for hand gesture recognition on skeletal data.* In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 106–113. IEEE, 2018. (Cited on page 14.)

[Diedrichsen & Kornysheva 2015] Jörn Diedrichsen and Katja Kornysheva. *Motor skill learning between selection and execution.* Trends in Cognitive Sciences, vol. 19, no. 4, pages 227–233, 2015. (Cited on page 62.)

[Dimitropoulos *et al.* 2018] Kosmas Dimitropoulos, Filareti Tsalakanidou, Spiros Nikolopoulos, Ioannis Kompatsiaris, Nikos Grammalidis, Sotiris Manitsaris, Bruce Denby, Lise Crevier-Buchman, Stephane Dupont, Vasileios Charisis*et al.* *A multimodal approach for the safeguarding and transmission of intangible cultural heritage: The case of i-Treasures.* IEEE Intelligent Systems, vol. 33, no. 6, pages 3–16, 2018. (Cited on page 61.)

[Duan *et al.* 2017] Yan Duan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel and Wojciech Zaremba. *One-shot imitation learning.* In Advances in neural information processing systems, pages 1087–1098, 2017. (Cited on page 64.)

[Duprey *et al.* 2017] Sonia Duprey, Alexandre Naaim, Florent Moissenet, Mickaël Begon and Laurence Cheze. *Kinematic models of the upper limb joints for multibody kinematics optimisation: An overview.* Journal of Biomechanics, vol. 62, pages 87–94, 2017. (Cited on page 12.)

[El Dine *et al.* 2018] Kamal Mohy El Dine, Jose Sanchez, Juan Antonio Corrales, Youcef Mezouar and Jean-Christophe Fauroux. *Force-torque sensor disturbance observer using deep learning.* In International Symposium on Experimental Robotics, pages 364–374. Springer, 2018. (Cited on page 42.)

[El-Shamouty *et al.* 2020] Mohamed El-Shamouty, Xinyang Wu, Shanqi Yang, Marcel Albus and Marco F Huber. *Towards Safe Human-Robot Collaboration Using Deep Reinforcement Learning.* In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 4899–4905. IEEE, 2020. (Cited on page 43.)

[El Zaatari *et al.* 2019] Shirine El Zaatari, Mohamed Marei, Weidong Li and Zahid Usman. *Cobot programming for collaborative industrial tasks: An overview.* Robotics and Autonomous Systems, vol. 116, pages 162–180, 2019. (Cited on page 40.)

[Faber *et al.* 2016] Gert S Faber, CC Chang, I Kingma, JT Dennerlein and JH Van Dieën. *Estimating 3D L5/S1 moments and ground reaction forces during trunk bending using a full-body ambulatory inertial motion capture system.* Journal of Biomechanics, vol. 49, no. 6, pages 904–912, 2016. (Cited on page 12.)

[Fang *et al.* 2017] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai and Cewu Lu. *RMPE: Regional Multi-person Pose Estimation.* In ICCV, 2017. (Cited on page 14.)

[Fardi *et al.* 2005] Basel Fardi, Ullrich Schuenert and Gerd Wanielik. *Shape and motion-based pedestrian detection in infrared images: a multi sensor approach.* In IEEE Proceedings. Intelligent Vehicles Symposium, 2005., pages 18–23. IEEE, 2005. (Cited on page 15.)

[Feichtenhofer *et al.* 2016] Christoph Feichtenhofer, Axel Pinz and Andrew Zisserman. *Convolutional two-stream network fusion for video action recognition.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1933–1941, 2016. (Cited on page 14.)

[Ferguson *et al.* 2014] Sam Ferguson, Emery Schubert and Catherine J Stevens. *Dynamic dance warping: Using dynamic time warping to compare dance move-*

ment performed under different conditions. In Proceedings of the 2014 International Workshop on Movement and Computing, pages 94–99, 2014. (Cited on page 65.)

[Fiebrink 2017] Rebecca Fiebrink. *Machine learning as meta-instrument: Human-machine partnerships shaping expressive instrumental creation.* In Musical instruments in the 21st century, pages 137–151. Springer, 2017. (Cited on page 80.)

[Finn *et al.* 2016] Chelsea Finn, Sergey Levine and Pieter Abbeel. *Guided cost learning: Deep inverse optimal control via policy optimization.* In International conference on machine learning, pages 49–58, 2016. (Cited on page 64.)

[Finn *et al.* 2017] Chelsea Finn, Pieter Abbeel and Sergey Levine. *Model-agnostic meta-learning for fast adaptation of deep networks.* arXiv preprint arXiv:1703.03400, 2017. (Cited on page 64.)

[Folgado *et al.* 2018] Duarte Folgado, Marília Barandas, Ricardo Matias, Rodrigo Martins, Miguel Carvalho and Hugo Gamboa. *Time alignment measurement for time series.* Pattern Recognition, vol. 81, pages 268–279, 2018. (Cited on page 65.)

[Françoise & Bevilacqua 2018] Jules Françoise and Frederic Bevilacqua. *Motion-sound mapping through interaction: An approach to user-centered design of auditory feedback using machine learning.* ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 8, no. 2, pages 1–30, 2018. (Cited on page 63.)

[Françoise 2015] Jules Françoise. *Motion-sound Mapping By Demonstration.* PhD thesis, 2015. (Cited on pages 13 and 82.)

[Frank *et al.* 2017] Malcolm Frank, Paul Roehrig and Ben Pring. What to do when machines do everything: How to get ahead in a world of ai, algorithms, bots, and big data. John Wiley & Sons, 2017. (Cited on pages 4 and 5.)

[Friberg & Sundberg 1999] Anders Friberg and Johan Sundberg. *Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners.* The Journal of the Acoustical Society of America, vol. 105, no. 3, pages 1469–1484, 1999. (Cited on page 82.)

[Gholipour & Arjmand 2016] Ali Gholipour and Navid Arjmand. *Artificial neural networks to predict 3D spinal posture in reaching and lifting activities; Applications in biomechanical models.* Journal of Biomechanics, vol. 49, no. 13, pages 2946–2952, 2016. (Cited on page 12.)

[Gildert *et al.* 2018] Naomi Gildert, Alan G. Millard, Andrew Pomfret and Jon Timmis. *The Need for Combining Implicit and Explicit Communication in Cooperative Robotic Systems.* Frontiers in Robotics and AI, vol. 5, no. 65, 2018. (Cited on page 42.)

[Gillian *et al.* 2011] Nicholas Gillian, Benjamin Knapp and Sile O'modhrain. *Recognition Of Multivariate Temporal Musical Gestures Using N-Dimensional Dynamic Time Warping.* In Nime, pages 337–342, 2011. (Cited on page 65.)

[Glushkova & Manitsaris 2015] Alina Glushkova and Sotiris Manitsaris. *Gesture Recognition Technologies for Gestural Know-how Management.* In Proceedings of the 7th International Conference on Computer Supported Education-Volume 1, pages 405–410, 2015. (Cited on pages 65 and 68.)

[Glushkova 2016] Alina Glushkova. *Gesture recognition technologies in managing movement skills. Sensori-motor feedback as a gamification mechanism.* PhD thesis, 2016. (Cited on page 60.)

[Gui *et al.* 2018] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan and José MF Moura. *Few-shot human motion prediction via meta-learning.* In Proceedings of the European Conference on Computer Vision (ECCV), pages 432–450, 2018. (Cited on page 64.)

[Güler *et al.* 2018] Rıza Alp Güler, Natalia Neverova and Iasonas Kokkinos. *Densepose: Dense human pose estimation in the wild.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7297–7306, 2018. (Cited on page 14.)

[Guo *et al.* 2018] Yijie Guo, Junhyuk Oh, Satinder Singh and Honglak Lee. *Generative adversarial self-imitation learning.* arXiv preprint arXiv:1812.00950, 2018. (Cited on page 64.)

[Han & Gold 2014] Jihyun Han and Nicolas Gold. *Lessons learned in exploring the Leap Motion™ sensor for gesture-based instrument design.* In Proceedings of the International Conference on New Interfaces for Musical Expression. Goldsmiths University of London, 2014. (Cited on page 83.)

[Hantrakul & Kaczmarek 2014] Lamtharn Hantrakul and Konrad Kaczmarek. *Implementations of the Leap Motion device in sound synthesis and interactive live performance.* In Proceedings of the 2014 International Workshop on Movement and Computing, pages 142–145, 2014. (Cited on page 83.)

[Harrison *et al.* 2011] Chris Harrison, Hrvoje Benko and Andrew D Wilson. *OmniTouch: wearable multitouch interaction everywhere.* In Proceedings of the 24th Annual ACM symposium on User Interface Software and Technology, pages 441–450, 2011. (Cited on page 83.)

[Hemery 2017] Edgar Hemery. *Modeling, recognition of finger gestures and upper-body movements for musical interaction design.* PhD thesis, 2017. (Cited on pages 80 and 91.)

[Hentout *et al.* 2019a] Abdelfetah Hentout, Mustapha Aouache, Abderraouf Maoudj and Isma Akli. *Human–robot interaction in industrial collaborative* *robotics: a literature review of the decade 2008–2017*. Advanced Robotics, vol. 33, no. 15-16, pages 764–799, 2019. (Cited on page 40.)

[Hentout *et al.* 2019b] Abdelfetah Hentout, Mustapha Aouache, Abderraouf Maoudj and Isma Akli. *Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017*. Advanced Robotics, vol. 33, no. 15-16, pages 764–799, 2019. (Cited on page 42.)

[Heo *et al.* 2019] Young Jin Heo, Dayeon Kim, Woongyong Lee, Hyoungkyun Kim, Jonghoon Park and Wan Kyun Chung. *Collision detection for industrial collaborative robots: a deep learning approach*. IEEE Robotics and Automation Letters, vol. 4, no. 2, pages 740–746, 2019. (Cited on page 42.)

[Ho & Ermon 2016] Jonathan Ho and Stefano Ermon. *Generative adversarial imitation learning*. arXiv preprint arXiv:1606.03476, 2016. (Cited on page 64.)

[Holden *et al.* 2016] Daniel Holden, Jun Saito and Taku Komura. *A deep learning framework for character motion synthesis and editing*. ACM Transactions on Graphics (TOG), vol. 35, no. 4, pages 1–11, 2016. (Cited on page 63.)

[Hospedales *et al.* 2020] Timothy Hospedales, Antreas Antoniou, Paul Micaelli and Amos Storkey. *Meta-learning in neural networks: A survey*. arXiv preprint arXiv:2004.05439, 2020. (Cited on page 64.)

[Jacob *et al.* 2015] Yannick Jacob, Sotiris Manitsaris, Fabien Moutarde, Gautam Lele and Laetitia Pradere. *Hand gesture recognition for driver vehicle interaction*. In IEEE Computer Society Workshop on Observing and Understanding Hands in Action (Hands 2015) of 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2015), 2015. (Cited on page 104.)

[Jaffe & Smith 1983] David A Jaffe and Julius O Smith. *Extensions of the Karplus-Strong plucked-string algorithm*. Computer Music Journal, vol. 7, no. 2, pages 56–69, 1983. (Cited on page 91.)

[Jégo *et al.* 2013] Jean-François Jégo, Alexis Paljic and Philippe Fuchs. *User-defined gestural interaction: A study on gesture memorization*. In 2013 IEEE Symposium on 3D User Interfaces (3DUI), pages 7–10. IEEE, 2013. (Cited on page 65.)

[Jordà *et al.* 2007] Sergi Jordà, Günter Geiger, Marcos Alonso and Martin Kaltenbrunner. *The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces*. In Proceedings of the 1st International Conference on Tangible and Embedded Interaction, pages 139–146, 2007. (Cited on page 83.)

[Keogh & Pazzani 2001] Eamonn J Keogh and Michael J Pazzani. *Derivative dynamic time warping*. In Proceedings of the 2001 SIAM International Conference on Data Mining, pages 1–11. SIAM, 2001. (Cited on page 65.)

[Kikui *et al.* 2018] Kosuke Kikui, Yuta Itoh, Makoto Yamada, Yuta Sugiura and Maki Sugimoto. *Intra-/inter-user adaptation framework for wearable gesture sensing device.* In Proceedings of the 2018 ACM International Symposium on Wearable Computers, pages 21–24, 2018. (Cited on page 63.)

[Kingma & Ba 2014] Diederik P Kingma and Jimmy Ba. *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014. (Cited on page 30.)

[Kitago & Krakauer 2013] Tomoko Kitago and John W Krakauer. *Motor learning principles for neurorehabilitation.* In Handbook of clinical neurology, volume 110, pages 93–103. Elsevier, 2013. (Cited on page 61.)

[Kober *et al.* 2013] Jens Kober, J Andrew Bagnell and Jan Peters. *Reinforcement learning in robotics: A survey.* The International Journal of Robotics Research, vol. 32, no. 11, pages 1238–1274, 2013. (Cited on page 64.)

[Kooij *et al.* 2019] Julian FP Kooij, Fabian Flohr, Ewoud AI Pool and Dariu M Gavrila. *Context-based path prediction for targets with switching dynamics.* International Journal of Computer Vision, vol. 127, no. 3, pages 239–262, 2019. (Cited on page 16.)

[Kopp *et al.* 2020] Tobias Kopp, Marco Baumgartner and Steffen Kinkel. *Success factors for introducing industrial human-robot interaction in practice: an empirically driven framework.* The International Journal of Advanced Manufacturing Technology, pages 1–20, 2020. (Cited on page 40.)

[Kucner *et al.* 2017] Tomasz Piotr Kucner, Martin Magnusson, Erik Schaffernicht, Victor Hernandez Bennetts and Achim J Lilienthal. *Enabling flow awareness for mobile robots in partially observable environments.* IEEE Robotics and Automation Letters, vol. 2, no. 2, pages 1093–1100, 2017. (Cited on page 16.)

[Lee & Kim 1999] Hyeon-Kyu Lee and Jin-Hyung Kim. *An HMM-based threshold model approach for gesture recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pages 961–973, 1999. (Cited on page 12.)

[Lee *et al.* 2017] Wonil Lee, Edmund Seto, Ken-Yu Lin and Giovanni C Migliaccio. *An evaluation of wearable sensors and their placements for analyzing construction worker's trunk posture in laboratory conditions.* Applied Ergonomics, vol. 65, pages 424–436, 2017. (Cited on page 17.)

[Leman 2010] Marc Leman. *Music, gesture, and the formation of embodied meaning.* In Musical Gestures, pages 138–165. Routledge, 2010. (Cited on page 82.)

[Li *et al.* 2019] Yong Li, Zihang He, Xiang Ye, Zuguo He and Kangrong Han. *Spatial temporal graph convolutional networks for skeleton-based dynamic hand* *gesture recognition.* EURASIP Journal on Image and Video Processing, vol. 2019, no. 1, page 78, 2019. (Cited on page 14.)

[Liu & Hao 2019] Zhiguang Liu and Jianhong Hao. *Intention Recognition in Physical Human-Robot Interaction Based on Radial Basis Function Neural Network.* Journal of Robotics, vol. 2019, 2019. (Cited on pages 42 and 43.)

[Lu & Chang 2012] Tung-Wu Lu and Chu-Fen Chang. *Biomechanics of human movement and its clinical applications.* The Kaohsiung Journal of Medical Sciences, vol. 28, pages S13–S25, 2012. (Cited on page 12.)

[Makovski 2018] Tal Makovski. *Meaning in learning: Contextual cueing relies on objects' visual features and not on objects' meaning.* Memory & Cognition, vol. 46, no. 1, pages 58–67, 2018. (Cited on page 62.)

[Makridakis *et al.* 2008] Spyros Makridakis, Steven C Wheelwright and Rob J Hyndman. Forecasting methods and applications. John Wiley & Sons, 2008. (Cited on page 31.)

[Manitsaris & Pekos 2008] Sotirios Manitsaris and Georgios Pekos. *Computer vision method for pianist's fingers information retrieval.* In Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services, pages 604–608, 2008. (Cited on page 89.)

[Manitsaris & Pekos 2009] Sotirios Manitsaris and Georgios Pekos. *Computer vision method in music interaction.* In 2009 First International Conference on Advances in Multimedia, pages 146–151. IEEE, 2009. (Cited on page 89.)

[Manitsaris *et al.* 2012] Sotiris Manitsaris, Bruce Denby, Florent Xavier, Jun Cai, Maureen Stone, Pierre Roussel and Gérard Dreyfus. *An open source speech synthesis module for a visual-speech recognition system.* In Acoustics 2012, 2012. (Cited on page 104.)

[Manitsaris *et al.* 2014a] Sotiris Manitsaris, Alina Glushkova, Frédéric Bevilacqua and Fabien Moutarde. *Capture, modeling, and recognition of expert technical gestures in wheel-throwing art of pottery.* Journal on Computing and Cultural Heritage (JOCCH), vol. 7, no. 2, pages 1–15, 2014. (Cited on page 62.)

[Manitsaris *et al.* 2014b] Sotiris Manitsaris, Alina Glushkova, Frédéric Bevilacqua and Fabien Moutarde. *Capture, modeling, and recognition of expert technical gestures in wheel-throwing art of pottery.* Journal on Computing and Cultural Heritage (JOCCH), vol. 7, no. 2, pages 1–15, 2014. (Cited on pages 68 and 69.)

[Manitsaris *et al.* 2015] Sotiris Manitsaris, Apostolos Tsagaris, Kosmas Dimitropoulos and Athanasios Manitsaris. *Finger musical gesture recognition in 3D space without any tangible instrument for performing arts.* International Journal of Arts and Technology, vol. 8, no. 1, pages 11–29, 2015. (Cited on page 89.)

[Manitsaris *et al.* 2016] Sotiris Manitsaris, Apostolos Tsagaris, Alina Glushkova, Fabien Moutarde and Frédéric Bevilacqua. *Fingers gestures early-recognition with a unified framework for RGB or depth camera.* In Proceedings of the 3rd International Symposium on Movement and Computing, pages 1–8, 2016. (Cited on page 89.)

[Manitsaris *et al.* 2020] Sotiris Manitsaris, Gavriela Senteri, Dimitrios Makrygiannis and Alina Glushkova. *Human movement representation on multivariate time series for recognition of professional gestures and forecasting their trajectories.* Frontiers in Robotics and AI, vol. 7, page 80, 2020. (Cited on page 63.)

[Manitsaris 2010] Sotiris Manitsaris. *Computer vision for the gesture recognition: gesture analysis and stochastic modelling in music interaction.* PhD thesis, 2010. (Cited on page 80.)

[Menychtas *et al.* 2019] Dimitrios Menychtas, Alina Glushkova and Sotirios Manitsaris. *Extracting the Inertia Properties of the Human Upper Body Using Computer Vision.* In International Conference on Computer Vision Systems, pages 596–603. Springer, 2019. (Cited on page 105.)

[Menychtas *et al.* 2020] Dimitrios Menychtas, Alina Glushkova and Sotiris Manitsaris. *Analyzing the Kinematic & Kinetic Contributions of the Human Upper Body's Joints for Ergonomics Assessment.* Journal of Ambient Intelligence and Humanized Computing, 2020. (Cited on pages 12 and 105.)

[Michalos *et al.* 2015] George Michalos, Sotiris Makris, Panagiota Tsarouchi, Toni Guasch, Dimitris Kontovrakis and George Chryssolouris. *Design considerations for safe human-robot collaborative workplaces.* Procedia CIrP, vol. 37, pages 248–253, 2015. (Cited on page 41.)

[Michalos *et al.* 2018] George Michalos, Niki Kousi, Panagiotis Karagiannis, Christos Gkournelos, Konstantinos Dimoulas, Spyridon Koukas, Konstantinos Mparis, Apostolis Papavasileiou and Sotiris Makris. *Seamless human robot collaborative assembly – An automotive case study.* Mechatronics, vol. 55, pages 194–211, 2018. (Cited on page 42.)

[Mínguez *et al.* 2018] Raúl Quintero Mínguez, Ignacio Parra Alonso, David Fernández-Llorca and Miguel Ángel Sotelo. *Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition.* IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 5, pages 1803–1814, 2018. (Cited on page 16.)

[Mitra & Acharya 2007] Sushmita Mitra and Tinku Acharya. *Gesture recognition: A survey.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 3, pages 311–324, 2007. (Cited on page 12.)

[Mohammadi Amin *et al.* 2020] Fatemeh Mohammadi Amin, Maryam Rezayati, Hans Wernher van de Venn and Hossein Karimpour. *A mixed-perception approach for safe human–robot collaboration in industrial automation.* Sensors, vol. 20, no. 21, page 6347, 2020. (Cited on page 43.)

[Mohammed *et al.* 2017] Abdullah Mohammed, Bernard Schmidt and Lihui Wang. *Active collision avoidance for human–robot collaboration driven by vision sensors.* International Journal of Computer Integrated Manufacturing, vol. 30, pages 970–980, 2017. (Cited on page 42.)

[Muller *et al.* 2020] Antoine Muller, Charles Pontonnier, Xavier Robert-Lachaine, Georges Dumont and André Plamondon. *Motion-based prediction of external forces and moments and back loading during manual material handling tasks.* Applied Ergonomics, vol. 82, page 102935, 2020. (Cited on page 12.)

[Nakra *et al.* 2009] Teresa Marrin Nakra, Yuri Ivanov, Paris Smaragdis and Christopher Ault. *The UBS Virtual Maestro: an Interactive Conducting System.* In NIME, pages 250–255, 2009. (Cited on page 83.)

[Newell 1991] Karl M Newell. *Motor skill acquisition.* Annual Review of Psychology, vol. 42, no. 1, pages 213–237, 1991. (Cited on page 62.)

[Olivas *et al.* 2019] Brenda Olivas, Alina Glushkova, Dimitrios Menychtas and Sotiris Manitsaris. *Designing a web-based automatic ergonomic assessment using motion data.* In Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pages 528–534, 2019. (Cited on page 105.)

[Olivas *et al.* 2020a] Brenda Olivas, Alina Glushkova and Sotiris Manitsaris. *Motion analysis for identification of overused body segments: the packaging task in industry 4.0.* Human Computer Interaction and Emerging Technologies: Adjunct Proceedings from, page 349, 2020. (Cited on page 105.)

[Olivas *et al.* 2020b] Brenda Elizabeth Olivas, Dimitrios Menychtas, Alina Glushkova and Sotiris Manitsaris. *Hidden Markov Modelling And Recognition Of Euler-Based Motion Patterns For Automatically Detecting Risks Factors From The European Assembly Worksheet.* In 2020 IEEE International Conference on Image Processing (ICIP), pages 3259–3263. IEEE, 2020. (Cited on page 105.)

[Papanagiotou *et al.* 2021] Dimitris Papanagiotou, Gavriela Senteri and Sotiris Manitsaris. *Egocentric gesture recognition using 3d convolutional neural networks for the spatio-temporal adaptation of collaborative robot, submitted.* Frontiers in Neurorobotics, 2021. (Cited on page 106.)

[Piaget 1976] Jean Piaget. *Piaget's theory.* In Piaget and his school, pages 11–23. Springer, 1976. (Cited on pages 62 and 68.)

[Pool *et al.* 2017] Ewoud AI Pool, Julian FP Kooij and Dariu M Gavrila. *Using road topology to improve cyclist path prediction.* In 2017 IEEE Intelligent Vehicles Symposium (IV), pages 289–296. IEEE, 2017. (Cited on page 16.)

[Pradere *et al.* 2019] Laetitia Pradere, Franck Guillemard, Gautam Lele, Fabien Moutarde, Yannick Jacob and Sotiris Manitsaris. *DISPOSITIF DE DETECTION DE GESTUELLES DE DOIGT (S) ET DE MAIN D'UN CONDUCTEUR DE VEHICULE POUR CONTROLER DES FONCTIONS*, 2019. (Cited on page 104.)

[Rabiner 1989] Lawrence R Rabiner. *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE, vol. 77, no. 2, pages 257–286, 1989. (Cited on pages 12 and 25.)

[Rad & Furlanello 2016] Nastaran Mohammadian Rad and Cesare Furlanello. *Applying deep learning to stereotypical motor movement detection in autism spectrum disorders.* In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pages 1235–1242. IEEE, 2016. (Cited on page 63.)

[Rauhala *et al.* 2008] Jukka Rauhala, Mikael Laurson, Vesa Välimäki, Heidi-Maria Lehtonen and Vesa Norilo. *A parametric piano synthesizer.* Computer Music Journal, vol. 32, no. 4, pages 17–30, 2008. (Cited on page 91.)

[Rigal 2003] Robert Rigal. Motricité humaine-tome 2: Fondements et applications pédagogiques, volume 2. PUQ, 2003. (Cited on page 65.)

[Rodgers 2010] Tara Rodgers. Pink noises: Women on electronic music and sound. Duke University Press, 2010. (Cited on page 82.)

[Rudenko *et al.* 2020] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila and Kai O Arras. *Human motion trajectory prediction: A survey.* The International Journal of Robotics Research, vol. 39, no. 8, pages 895–935, 2020. (Cited on page 17.)

[Ruffaldi *et al.* 2009] Emanuele Ruffaldi, Alessandro Filippeschi, Antonio Frisoli, Oscar Sandoval, Carlo Alberto Avizzano and Massimo Bergamasco. *Vibrotactile perception assessment for a rowing training system.* In World Haptics 2009-Third Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pages 350–355. IEEE, 2009. (Cited on page 66.)

[Safeea *et al.* 2019] Mohammad Safeea, Pedro Neto and Richard Bearee. *On-line collision avoidance for collaborative robot manipulators by adjusting off-line generated paths: An industrial use case.* Robotics and Autonomous Systems, vol. 119, pages 278–288, 2019. (Cited on page 42.)

[Schank 1997] Roger Schank. Virtual learning. a revolutionary approach to building a highly skilled workforce. ERIC, 1997. (Cited on page 62.)

[Schmidt & Wrisberg 2008] Richard A Schmidt and Craig A Wrisberg. Motor learning and performance: A situation-based learning approach. Human Kinetics, 2008. (Cited on page 65.)

[Schmidtler *et al.* 2015] Jonas Schmidtler, Verena Knott, Christin Hölzel and Klaus Bengler. *Human Centered Assistance Applications for the working environment of the future*. Occupational Ergonomics, vol. 12, no. 3, pages 83–95, 2015. (Cited on page 41.)

[Schneider & Gavrila 2013] Nicolas Schneider and Dariu M Gavrila. *Pedestrian path prediction with recursive bayesian filters: A comparative study*. In German Conference on Pattern Recognition, pages 174–183. Springer, 2013. (Cited on page 15.)

[Schwarz *et al.* 2012] Loren Arthur Schwarz, Artashes Mkhitaryan, Diana Mateus and Nassir Navab. *Human skeleton tracking from depth data using geodesic distances and optical flow*. Image and Vision Computing, vol. 30, no. 3, pages 217–226, 2012. (Cited on page 45.)

[Shahroudy *et al.* 2016] Amir Shahroudy, Jun Liu, Tian-Tsong Ng and Gang Wang. *Ntu rgb+ d: A large scale dataset for 3d human activity analysis*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1010–1019, 2016. (Cited on page 14.)

[Sharkawy *et al.* 2020a] Abdel-Nasser Sharkawy, Panagiotis N Koustoumpardis and Nikos Aspragathos. *Human–robot collisions detection for safe human–robot interaction using one multi-input–output neural network*. Soft Computing, vol. 24, no. 9, pages 6687–6719, 2020. (Cited on page 42.)

[Sharkawy *et al.* 2020b] Abdel-Nasser Sharkawy, Panagiotis N Koustoumpardis and Nikos Aspragathos. *Neural network design for manipulator collision detection based only on the joint position sensors*. Robotica, vol. 38, no. 10, pages 1737–1755, 2020. (Cited on page 42.)

[Shmuelof *et al.* 2012] Lior Shmuelof, John W Krakauer and Pietro Mazzoni. *How is a motor skill learned? Change and invariance at the levels of task success and trajectory control*. Journal of Neurophysiology, vol. 108, no. 2, pages 578–594, 2012. (Cited on page 62.)

[Shojaei *et al.* 2016] Iman Shojaei, Milad Vazirian, Emily Croft, Maury A Nussbaum and Babak Bazrgari. *Age related differences in mechanical demands imposed on the lower back by manual material handling tasks*. Journal of Biomechanics, vol. 49, no. 6, pages 896–903, 2016. (Cited on page 12.)

[Sigrist *et al.* 2013] Roland Sigrist, Georg Rauter, Robert Riener and Peter Wolf. *Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review.* Psychonomic Bulletin & Review, vol. 20, no. 1, pages 21–53, 2013. (Cited on pages 65, 66 and 67.)

[Song *et al.* 2016] Sibo Song, Vijay Chandrasekhar, Bappaditya Mandal, Liyuan Li, Joo-Hwee Lim, Giduthuri Sateesh Babu, Phyo Phyo San and Ngai-Man Cheung. *Multimodal multi-stream deep learning for egocentric activity recognition.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 24–31, 2016. (Cited on page 15.)

[Soomro *et al.* 2012] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah. *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild.* CoRR, vol. abs/1212.0402, 2012. (Cited on page 30.)

[Srikanth *et al.* 2019] Shashank Srikanth, Junaid Ahmed Ansari, R Karnik Ram, Sarthak Sharma, J Krishna Murthy and K Madhava Krishna. *Infer: Intermediate representations for future prediction.* In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 942–949. IEEE, 2019. (Cited on pages 16 and 17.)

[Sun *et al.* 2018] Li Sun, Zhi Yan, Sergi Molina Mellado, Marc Hanheide and Tom Duckett. *3DOF pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data.* In 2018 IEEE International Conference on Robotics and Automation (ICRA), pages 5942–5948. IEEE, 2018. (Cited on page 16.)

[Tao & Liu 2013] Chongben Tao and Guodong Liu. *A multilayer hidden Markov models-based method for human-robot interaction.* Mathematical Problems in Engineering, vol. 2013, 2013. (Cited on page 42.)

[Taralle *et al.* 2015a] Florent Taralle, Alexis Paljic, Sotiris Manitsaris, Jordane Grenier and Christophe Guettier. *A Consensual and Non-ambiguous Set of Gestures to Interact with UAV in Infantrymen.* In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, pages 797–803, 2015. (Cited on page 105.)

[Taralle *et al.* 2015b] Florent Taralle, Alexis Paljic, Sotiris Manitsaris, Jordane Grenier and Christophe Guettier. *Is symbolic gestural interaction better for the visual attention?* Procedia Manufacturing, vol. 3, pages 1060–1065, 2015. (Cited on page 105.)

[Taralle 2016] Florent Taralle. *Guidage Gestuel pour des Robots Mobiles.* PhD thesis, Paris Sciences et Lettres, 2016. (Cited on page 105.)

[Tchernichovski *et al.* 2000] Ofer Tchernichovski, Fernando Nottebohm, Ching Elizabeth Ho, Bijan Pesaran and Partha Pratim Mitra. *A procedure for an au-*

tomated measurement of song similarity. Animal Behaviour, vol. 59, no. 6, pages 1167–1176, 2000. (Cited on page 99.)

[Tilmanne 2013] Joëlle Tilmanne. *Data-driven Stylistic Humanlike Walk Synthesis.* PhD thesis, PhD Dissertation, University of Mons, 2013.(Cited on pages 28 and 55.), 2013. (Cited on pages 12 and 63.)

[Tormoen *et al.* 2014] Daniel Tormoen, Florian Thalmann and Guerino Mazzola. *The Composing Hand: Musical Creation with Leap Motion and the BigBang Rubette.* In Proceedings of the International Conference on New Interfaces for Musical Expression, pages 207–212, 2014. (Cited on page 83.)

[Tran *et al.* 2015a] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri. *Learning Spatiotemporal Features with 3D Convolutional Networks.* In 2015 IEEE International Conference on Computer Vision (ICCV), pages 4489–4497, 2015. (Cited on page 30.)

[Tran *et al.* 2015b] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani and Manohar Paluri. *Learning spatiotemporal features with 3d convolutional networks.* In Proceedings of the IEEE International Conference on Computer Vision, pages 4489–4497, 2015. (Cited on page 14.)

[Tsagaris *et al.* 2011] Apostolos Tsagaris, Sotiris Manitsaris, Kosmas Dimitropoulos and Athanasios Manitsaris. *Intelligent invariance techniques for music gesture recognition based on skin modelling.* In 2011 IEEE 12th International Symposium on Computational Intelligence and Informatics (CINTI), pages 219–223. IEEE, 2011. (Cited on page 89.)

[Volioti *et al.* 2014] Christina Volioti, Sotiris Manitsaris and Athanasios Manitsaris. *Offline statistical analysis of gestural skills in pottery interaction.* In Proceedings of the 2014 International Workshop on Movement and Computing, pages 172–173, 2014. (Cited on pages 73 and 92.)

[Volioti *et al.* 2015] Christina Volioti, Edgar Hemery, Sotiris Manitsaris, Vicky Teskouropoulou, Erdal Yilmaz, Fabien Moutarde and Athanasios Manitsaris. *Music gestural skills development engaging teachers, learners and expert performers.* Procedia Manufacturing, vol. 3, pages 1543–1550, 2015. (Cited on pages 92 and 95.)

[Volioti *et al.* 2016a] Christina Volioti, Stelios Hadjidimitriou, Sotiris Manitsaris, Leontios Hadjileontiadis, Vasileios Charisis and Athanasios Manitsaris. *On mapping emotional states and implicit gestures to sonification output from the'Intangible Musical Instrument'.* In Proceedings of the 3rd International Symposium on Movement and Computing, pages 1–5, 2016. (Cited on page 92.)

[Volioti *et al.* 2016b] Christina Volioti, Sotiris Manitsaris, Eleni Katsouli and Athanasios Manitsaris. *x2Gesture: how machines could learn expressive* *gesture variations of expert musicians.* In Proceedings of the International Conference on New Interfaces for Musical Expression, pages 310–315, 2016. (Cited on page 92.)

[Volioti *et al.* 2018] Christina Volioti, Sotiris Manitsaris, Edgar Hemery, Stelios Hadjidimitriou, Vasileios Charisis, Leontios Hadjileontiadis, Eleni Katsouli, Fabien Moutarde and Athanasios Manitsaris. *A natural user interface for gestural expression and emotional elicitation to access the musical intangible cultural heritage.* Journal on Computing and Cultural Heritage (JOCCH), vol. 11, no. 2, pages 1–20, 2018. (Cited on page 92.)

[Volioti 2016] Christina Volioti. *Machine learning in sonification of expressive gesture with the use of stochastic models.* PhD thesis, 2016. (Cited on pages 81 and 94.)

[Wickens 2008] Christopher D Wickens. *Multiple resources and mental workload.* Human Factors, vol. 50, no. 3, pages 449–455, 2008. (Cited on page 66.)

[Wolpert *et al.* 2011] Daniel M Wolpert, Jörn Diedrichsen and J Randall Flanagan. *Principles of sensorimotor learning.* Nature Reviews Neuroscience, vol. 12, no. 12, pages 739–751, 2011. (Cited on pages 62 and 66.)

[Xue *et al.* 2018] Hao Xue, Du Q Huynh and Mark Reynolds. *Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction.* In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1186–1194. IEEE, 2018. (Cited on page 16.)

[Yan *et al.* 2018] Sijie Yan, Yuanjun Xiong and Dahua Lin. *Spatial temporal graph convolutional networks for skeleton-based action recognition.* arXiv preprint arXiv:1801.07455, 2018. (Cited on page 14.)

[Yu *et al.* 2018] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel and Sergey Levine. *One-shot imitation from observing humans via domain-adaptive meta-learning.* arXiv preprint arXiv:1802.01557, 2018. (Cited on page 64.)

[Zandt-Escobar *et al.* 2014] Van Zandt-Escobar, Baptiste Caramiaux, Atau Tanaka*et al.* *Piaf: A tool for augmented piano performance using gesture variation following.* 2014. (Cited on page 92.)

[Zatsiorsky 2008] Vladimir Zatsiorsky. Biomechanics in sport: performance enhancement and injury prevention, volume 9. John Wiley & Sons, 2008. (Cited on page 12.)

[Zernetsch *et al.* 2016] Stefan Zernetsch, Sascha Kohnen, Michael Goldhammer, Konrad Doll and Bernhard Sick. *Trajectory prediction of cyclists using a physical model and an artificial neural network.* In 2016 IEEE Intelligent Vehicles Symposium (IV), pages 833–838. IEEE, 2016. (Cited on page 16.)

[Zhao & Badler 2001] Liwei Zhao and Norman I Badler. *Synthesis and acquisition of laban movement analysis qualitative parameters for communicative gestures.* Technical Reports (CIS), page 116, 2001. (Cited on page 81.)

[Zhu *et al.* 2018] Yuke Zhu, Ziyu Wang, Josh Merel, Andrei Rusu, Tom Erez, Serkan Cabi, Saran Tunyasuvunakool, János Kramár, Raia Hadsell, Nando de Freitas*et al. Reinforcement and imitation learning for diverse visuomotor skills.* arXiv preprint arXiv:1802.09564, 2018. (Cited on page 64.)