Survey on AI-Based Multimodal Methods for Emotion Detection

Catherine Maréchal, Dariusz Mikolajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, Corinne Ancourt, Katarzyna Wegrzyn-Wolska

Survey on AI-Based Multimodal Methods for Emotion Detection

Catherine Marechal[1], Dariusz Mikołajewski[2,3], Krzysztof Tyburek[2], Piotr Prokopowicz[2], Lamine Bougueroua[1], Corinne Ancourt[4], and Katarzyna Węgrzyn-Wolska[1(✉)]

[1] Allianstic Research Laboratory, Efrei Paris, Villejuif, France
{catherine.marechal, lamine.bougueroua, katarzyna.wegrzyn-wolska}@efrei.fr
[2] Institute of Mechanics and Applied Computer Science, Kazimierz Wielki University, Bydgoszcz, Poland
{dmikolaj, krzysiekkt, piotrek}@ukw.edu.pl
[3] Neurocognitive Laboratory, Centre for Modern Interdisciplinary Technologies, Nicolaus Copernicus University, Toruń, Poland
darek.mikolajewski@wp.pl
[4] MINES ParisTech, PSL University, CRI, Paris, France
corinne.ancourt@mines-paristech.fr

Abstract. Automatic emotion recognition constitutes one of the great challenges providing new tools for more objective and quicker diagnosis, communication and research. Quick and accurate emotion recognition may increase possibilities of computers, robots, and integrated environments to recognize human emotions, and response accordingly to them a social rules. The purpose of this paper is to investigate the possibility of automated emotion representation, recognition and prediction its state-of-the-art and main directions for further research. We focus on the impact of emotion analysis and state of the arts of multimodal emotion detection. We present existing works, possibilities and existing methods to analyze emotion in text, sound, image, video and physiological signals. We also emphasize the most important features for all available emotion recognition modes. Finally, we present the available platform and outlines the existing projects, which deal with multimodal emotion analysis.

Keywords: Affective computing · Emotion detection · Automatic data processing · Data collection · Expressed emotion · Big data · Artificial intelligence

1 Introduction

Affective Computing (AC) has been a popular area of research for several years. Many research has been conducted to enable machines detect and understand human affective states, such as emotions, interests and the behavior. It attempts to bridge the communication gap between human users and computers with "soulless" and "emotionless"

J. Kołodziej and H. González-Vélez (Eds.): cHiPSet, LNCS 11400, pp. 307–324, 2019.
https://doi.org/10.1007/978-3-030-16272-6_11

feeling. The inability of today's systems to recognize, express and feel emotions limits their ability to act intelligently and interact naturally with humans.

To become more user-friendly and effective, systems need to become sensitive to human emotions. Verbal and nonverbal information is very important because it complements the verbal message and provides a better interpretation of the message. It is very useful to design and develop systems that can measure the emotional state of a person based on, for example, gestures (body movements and postures), facial expression, acoustic characteristics and emotions expressed in the text. In the practical case, body signals and facial expressions recorded in real-time by sensors and cameras can be associated with predetermined emotions.

It is interesting to merge the multimodal information retrieved by these devices with information from analysis of emotions and intensity in text. For example, the ultimate goal may be to improve the system's decisions so that they can react accordingly to recognized emotions, which will allow better human-machine interaction.

Conventional wisdom says that 70–90% of communication between humans is nonverbal. The studies conducted by Albert Mehrabian in 1967 established the 7%–38%–55% rule, also known as the "3V rule": 7% of the communication is verbal, 38% of the communication is vocal and 55% of the communication is visual [1]. This justifies the interest and importance of nonverbal communication.

The first study in the field of emotion detection was born during the sixties/seventies. The most prominent example is that of mood rings [2]. The principle is simple; rings contain thermotropic liquid crystals that react with body temperature. When a person has stressed, his mood ring take on a darker color.

The scientific publications of Rosalind Picard (MIT) have introduced a great progress in this field since the nineties [3, 4]. He is one of the pioneers of affective computing. In his book "Affective Computing", Picard proposed that emotion can be modeled using the nonlinear sigmoid function. Over the last 20 years, the development of technology has allowed the implementation of relatively good system market and efficient such as ambient intelligence (AMI), virtual reality (VR) and augmented reality (AR).

Nowadays, in the automotive field for example, an on-board computer that is able to detect confusion, interest or fatigue can increase safety. The AutoEmotive (MIT Media Lab) is a prototype equipped with sensors and a camera placed on the steering wheel [5]. This vehicle measures the level of stress and fatigue of the driver. When the need arises, he puts a background music, changes the temperature and light in the vehicle interior, or still proposes to follow a less stressful journey.

We have structured the remainder of the paper as follows. Section 2 describes in general existing works dealing with the emotion detection and multimodal emotion analysis problem. Section 3 presents the basic modalities and methods used for emotion analysis. We have presented the related existing projects dealt with this problem in Sect. 4. We summarize the survey and introduce some directions for future research in Sect. 5.

2 Multimodal Emotion Analysis

Traditional emotion recognition and classification associate the "value" to at least several basic emotions such as happiness, joy, neutrality, surprise, sadness, disgust, anger, or fear.

Multimodal emotion analysis becomes very popular research domain. Started from the classic language-based definition of sentiment analysis was extended to a multimodal with other relevant modalities (video, audio, sensor data, etc.). Different techniques and methods are combined to achieve it; very often, they are based on big data methods; semantic rules and machine learning.

A novel multimodal implicit emotion recognition system can be built upon an AI-based model designed to extract information on the emotion from different devices. To feed such a model, a video data captured by the camera embedded in the user's device (laptop, desktop, tablet, etc.), an audio signals collected from microphones embedded in mobile devices, and motion signals generated by sensors in wearable devices can be used.

Several multimodal datasets include sentiment annotations. Zadeh et al. introduced the first multimodal dataset (MOSI) with opinion-level sentiment intensity annotations and studying the prototypical interaction patterns between facial gestures and spoken words when inferring sentiment intensity. They proposed a new computational representation, called multimodal dictionary, based on a language-gesture study and evaluated the proposed approach in a speaker-independent model for sentiment intensity prediction [6]. For other examples of data sets we can cite ICT-MMMO [7] and MOUD [8] datasets.

One of the most challenges in multimodal emotion analysis is to model the interactions between language, visual and acoustic behaviors that change the observation of the expressed emotion (named the inter-modality dynamics). A second challenge in multimodal emotion analysis (named intra-modality dynamics) is to efficiently explore emotion, not only on one but on highly expressive nature modality (ex. spoken language where proper language structure is often ignored, video and acoustic modalities which are expressed through both space and time. To solve this problem, Zadeh introduced the Tensor Fusion Network (TFN), which combine the intra-modality and inter-modality models. Inter-modality dynamics is modeled with a new multimodal fusion approach, named Tensor Fusion, which explicitly aggregates unimodal, bimodal and trimodal interactions. Intra-modality dynamics is modeled through three Modality Embedding Subnetworks, for language, visual and acoustic modalities, respectively [9]. Interesting work is realized by Poria et al., who developed a LTSM-based network to extract contextual features from the video for multimodal sentiment analysis [10].

Additionally, they presented a multimodal sentiment analysis framework, which includes sets of relevant features for text and visual data, as well as a simple technique for fusing the features extracted from different modalities [11].

The previous presented methods combine and analyze data from various types of modality, the basis for all these analyzes is appropriate detection of emotion on each modality input. At the same time, we must point out that for each modality individually

we can also perform emotion analysis. In the next chapter, we will describe the methods of emotion analysis for each modality separately.

3 Basic Modalities for Emotion Analysis

3.1 Emotion in Text

Opinion Mining (OM) and Sentiments Analysis (SA) consists in identifying orientation or intensity of sentiments and opinion in pieces of texts (blogs, forums, user comments, review websites, community websites, etc.). It enables determining whether a sentence or a document expresses positive, negative or neutral sentiment towards some object or more. In addition, it allows for classification of opinions/sentiments according to intensity degrees.

Definition. An opinion is a quadruple (*O, F, H, S*) where *O* is a target object, *F* is a set of features of the object *O*, *H* is a set of opinion's holders, *S* is the set of sentiment/opinion value of the opinion's holder on feature $f_i \in F$ of the object *O*.

According to Liu [12] "sentiment analysis is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions toward entities such as products, services, organizations, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc."

Sentiments analysis is a complex technique. Sentiments and opinions can often be expressed in a subtle manner that creates difficulty in the identification of their emotional value. Moreover, opinions and sentiments are highly sensitive to the context and the field in which they are used: the same string might be positive in one context and negative in another. In addition, on the Internet, everyone uses his or her own style and vocabulary, what adds extra difficulty to the task. It is not yet possible to find out an ideal case to marking the opinion in a text written by different users, because the text does not follow the rules. Therefore, it is impossible to schedule every possible case. Moreover, very often the same phrase can be considered as positive for one person and negative for another one.

Text analysis and social network analysis have gained importance with growing interest in Big Data. Both deal with large amounts of data, largely unstructured, and the Big Data benefit comes from the application of these two methods of data analysis. We will begin by examining some of the ways in which text analysis can be applied to sentiment analysis before moving on to social network analysis.

Text Analysis

Sentiment Analysis (SA) [13] is a computational study of how opinions, attitudes, emoticons and perspectives are expressed in language. Sentiment Detection, or in its simplified form – Polarity Classification, is a tedious and complex task. Contextual changes of polarity indicating words, such as negation, sarcasm as well as weak syntactic structures make it troublesome for both machines and humans to safely determine polarity of messages.

Sentiment analysis methods involve building a system to collect and categorize opinions about a product. This consists in examining natural language conversations happening around a certain product for tracking the mood of the public. The analysis is performed on large collections of texts, including web pages, online news, Internet discussion groups, online reviews, web blogs, and social media. Opinion Mining aims to determine polarity and intensity of a given text, i.e., whether it is positive, negative, or neutral and to what extent. To classify the intensity of opinions, we can use methods introduced in [14–16].

Social Networks

Social Networks are indisputably popular nowadays and show no sign of slowdown. According to the Kepios study [17], the number of active users of social networks increased by 13% in 2017 to reach 3.3 billion users in April 2018. For example, Facebook attracts more than 2.2 billion users a month. Penetrating ever more aspects of our daily life, they become not only a considerable threat to our privacy, but also an encompassing tool for analyzing opinions, habits, trends and some would even say – thoughts.

In the current growth of artificial intelligence, machine learning and natural language processing, driven by new technological possibilities, it is possible to automate the analysis of vast amounts of publicly published data.

Text Mining and Social Network Analysis have become a necessity for analyzing not only information but also the connections across them. The main objective is to identify the necessary information as efficiently as possible, finding the relationships between available information by applying algorithmic, statistical, and data management methods on the knowledge. The automation of sentiment detection on these social networks has gained attention for various purposes [18–20].

Twitter is a social network that allows the user to freely publish short messages, called Tweets via the Internet, instant messaging or SMS. These messages are limited to 140 characters (more exactly, NFC normalized code points [21]). With about 330 million monthly active users (as of 2018, Twitter Inc.), Twitter is a leading social network, which is known for its ease of use for mobile devices (90% of users access the social network via mobile devices). Twitter known by the diversity of content, as well as its comprehensive list of APIs offered to developers. With an average of 500 million messages sent per day, the platform seems ideal for tracking opinions on various subjects. Furthermore, the very short format messages facilitate classification since short messages rarely discuss more than one topic.

However, automated interpretation is complicated by embedded links, abbreviations and misspellings. Facing these challenges is becoming increasingly important for Economic and Market Intelligence in order to successfully recognize trends and threats.

The frequent expression of negative emotion words on social media has been linked to depression. The aim of [22] was to report on the associations between depression severity and the variability (time-unstructured) and instability (time-structured) in emotion word expression on Facebook and Twitter across status updates. Several works on depression have emerged. They are based on social networks: Twitter [23, 24] and Facebook [25, 26].

Several authors have been interested in the use of emoticons to complete the sentiment analysis. Authors in [27] utilize Twitter API to get training data that contain emoticons like :) and :(. They use these emoticons as noisy labels. Tweets with :) are thought to be positive training data and tweets with :(are thought to be negative training data. In [28], authors present the ESLAM (Emoticon Smoothed LAnguage Models) which combine fully supervised methods and distantly supervised methods. Although many TSA (Twitter Sentiment Analysis) methods have been presented. The authors in [29] explored the influence of emoticons on TSA.

3.2 Emotion Detection in the Sound

Automatic emotion recognition based on utterance level prosodic features may play an important role within speaker-independent emotion recognition [30]. The recognition of emotions based on the voice has been studied for decades [31–34]. However, most of the work-concerned data collected in a controlled environment in which the data are clean without significant noise and directly well segmented. In addition, the majority of such a system are speech-oriented. In the real world, the process is much more complex. There are many factors such as background noise and not speech voice like a laugh, a whimper, a cry, a sigh, etc., which greatly aggravate the results obtained in a controlled environment. These factors will make the real emotion recognition trained on the data from the controlled environment unsuccessful.

The author in [35] focused on mono-modal systems with speech as only input channel. He proposed FAU Aibo Emotion Corpus, which is a speech corpus with naturally occurring emotion-related states. FAU Aibo is a corpus of spontaneous, emotionally colored speech of German children at the age of 10 to 13 years interacting with the Sony Robot Aibo. Eleven emotion-related states are labeled on the word level. Best results have been obtained on the chunk level where a classwise averaged recognition rate of almost 70% for the 4-class problem anger, emphatic, neutral and motherese has been achieved.

Voice-based emotion recognition system relying on audio input has low requirements for hardware. Especially the recent emerging of speech-based artificial assistant, e.g. Google Home, Amazon Alexa, etc., provides the ready-to-employ platform for voice-based emotion recognition system.

Audio Analysis: Background

Nowadays human speech recognition (HSR) and automatic speech recognition (ASR) systems have a very wide application. This aspect is also referred to the recognition of emotion state. Emotion recognition by research human speech is connected with two research areas. First is related to synthetic approach, which allows generating artificial speech samples filled specific emotions. The second issue concerns machine recognition of the speaker's emotions. In order to possible, a machine (IT system) has to learn human emotions expressed out of speech like speech intonation, articulation, etc. By the recognizing emotions from human speech, we can notice which kind of emotion is dominant during conversation. Possible to recognize emotions state like sadness, fear, embarrassment, terror, etc. Connected with this, machines could help people in making right decisions by recognizing emotions, especially in irrational

situations where decisions have to be made faster than a rational performing mind. Sometimes it is valuable to artificially influence mental and emotional states to get a better individual performance in stress-related occupations and prevent mental disorders from happening [36]. Recent research has shown that under certain circumstances multimodal emotion recognition is possible even in real time [37].

Feature Extraction

Sound signals (including human speech) is one of the main mediums of communication and it can be processed to recognize the speaker or even emotion. This diagnosis is possible through signal decomposition and time-frequency analysis. During this analysis, key physical features are found which are clearly able to describe the emotional background of the speaker. Analyzing speech as a sound and not the meaning of spoken words is possible eliminate the language barrier by focusing only on the emotional message. This can be obtained by calculating the values of descriptors be able to describe such features as the ratio of amplitudes in particular parts of the signal, the shape of the signal waveform, the frequency distribution, etc. The basic principle behind emotion recognition lies with analyzing the acoustic difference that occurs when talking about the same thing under different emotional situations [38].

The accurate selection of descriptors and their combination, allows to determine which emotions are dominant in conversation - that is, their worth is very important for classification. It is therefore necessary to index the speech signal to evaluate the speaker's emotions. Acquiring features is possible by investigating time domain and frequency domain speech signals. This way, it is practicable to obtain feature vector which can be able to automatic objects classification - it means, automatic classification of emotion state too. During research is necessary concentrate on insightful research frequency domain, not forgetting the time domain descriptors either. There are some physical features applied for indexing speech, like: spectrum irregularity, wide and narrow band spectrograms, speech signals filtering and processing, enhancement and manipulation of specific frequency regions, segmentation and labeling of words, syllables and individual phonemes [37]. Moreover, the Mel-Frequency Cepstral Coefficients (MFCC) is widely used in speech classification experiments due to its good performance. It extracts and represents features of speech signal - including kinds of emotions. The Mel-Cepstral takes short-time spectral shape with important data about the quality of voice and production effects. To calculate these coefficients the inverse cosine transform of decimal logarithm of the Mel filtered short-term spectrum of energy must be done. The purpose of improving results of experiment and searching effective spectral feature vector, constant spectral resolution is used [39]. For the reduction of leakage effect, the Hamming window is implemented. This is necessary for increasing the efficiency of frequency in human speech [38].

There is MPEG 7 standard, which gives many descriptors definitions for the physical features of sound (including human speech). These descriptors are defined on the base of analysis of digital signals and index of most important their factors. The MPEG 7 Audio standard contains descriptors and description schemes that can be divided into two classes: generic low-level tools and application-specific tools [40]. MPEG 7 descriptors are being very helpful for assessment features of human speech in

a general sense. They can be used to assess the emotions because each emotional state expressed in speech contains specific numeric values of MPEG 7 descriptors.

After the extraction of features, the machine learning part of the generic system should analyze features and find statistical relationships between particular features and emotional states. This part of the system is also called "classifier". Most commonly used classification algorithms are Artificial Neural Networks (ANN), k-Nearest Neighbor (k-NN) and Support Vector Machines (SVM), decision trees. Furthermore, probabilistic models such as the Gaussian mixture model (GMM) or stochastic models such as Hidden Markov Model (HMM) can be applied [36].

Accuracy
Emotion analysis of speech is possible; however, it highly depends of the language. Automatic and universal emotion analysis is very challenging. Analyze of emotion is more efficient when perform for one dedicated language. Study by Chaspari et al. showed that emotion classification in speech (Greek language) achieved accuracy up to 75.15% [41]. Similar study by Arruti et al. showed mean accuracy of 80.05% emotion recognition rate in Basque and a 74.82% in Spanish [42].

3.3 Emotion in Image and Video

Nonverbal behavior constitutes useful means of communication in addition to spoken language. It allows to communicate even complex social concepts, in various contexts, thus may be regarded not only supplementary, but also as basic way of emotion recognition. It is possible to easily reading the affects, emotions and intentions from face expressions and gestures. Indeed, Ekman (one of the pioneers in the study of facial expressions and emotions) [43] identifies at least six characteristics from posed facial actions that enable emotion recognition: morphology, symmetry, duration, speed of onset, coordination of apexes and ballistic trajectory. They are common to all humans confirming Darwin's evolutionary thesis. Therefore, an emotional recognition tools based on facial video is universal.

However, automatic and either - automatic semi recognition of the meaning of face expressions and gesture still constitutes true challenge. We will focus on facial expressions.

3.4 Automatic Facial Emotion

The human face, as a modality for emotion detection takes a dominant position in the study of affect. It is possible to register the facial information automatically in real-time, without requiring any specialized equipment except a simple video camera. Thus, facial expressions as a noninvasive method are used in behavioral science and in clinical. Therefore, automatic recognition of facial expressions is an important component for emotion detection and modeling the natural human-machine interfaces.

The face is equipped with a large number of independent muscles, which can be activated at different levels of intensity. Consequently, the face is capable of generating a high number of complex dynamic facial expression patterns. There are three main dimensions of facial variation: morphology, complexion and dynamics. As first two

dimensions, morphology and complexion, deals with static dimensions, there are very useful for facial recognition but their role for emotion analyze is not significant. The third one, the dynamics, play the most important role for emotion face analysis [44]. Although mapping of facial expressions to affective states is possible, the automatic recognizing of humans' emotion from the facial expressions, without effort or delay is still a challenge.

Automatic detection of emotions from facial expressions are not simple and their interpretation is largely context-driven. To reduce the complexity of automatic affective inference, measurement and interpretation of facial expressions, Ekman and Friesen developed in 1978 special system for objectively measuring facial movement; the Facial Action Coding System (FACS) [45]. FACS, based on a system originally developed by a Swedish anatomist named Hjortsjö [46] became the standard for identifying any movement of the face. Later, Ekman and Sejnowski studied also computer based facial measurements [47].

FACS is a common standard to systematically categorize and to index the physical expression of emotions. The basic values of FACE are Action Units (*AUs*). *AUs* are the fundamental actions of individual muscles or groups of muscles. The *AUs* are grouped in several categories and identified by a number: main codes, head movement codes, eye movement codes, visibility codes, and gross behavior codes. The intensities of *AU*s are annotated in five categories, by appending letters A–E to the *AUs* (A for minimal intensity-trace, B for slight intensity, C for marked, D for severe, E for maximum intensity). For example, *AU1A* signifies the weakest trace of *AU1* and *AU1E* is the maximum intensity possible of *AU1* for the individual person.

The eyes and mouth have high importance to emotion detection; therefore, to successfully recognize an emotion, the observations mostly rely on the eye and mouth regions. Furthermore, the actions of eyes and mouth allowed grouping the expressions in a continuous space, ranging from sadness and fear (reliance on the eyes) to disgust and happiness (mouth). Combining these observations with facial *AUs* increase knowledge about the areas involved in displaying each emotion. For example, "happy" denoted by *AU6* + *AU12,* comprises *AU6* (Cheek Raiser) and *AU12* (Lip Corner Puller), whereas 'sad' (*AI1* + *AU4* + *AU15*) comprises *AU1* (Inner Brow Raiser), *AU4* (Brow Lowerer) and *AU15* (Lip Corner Depressor).

The computer algorithm for facial coding extracts the main features of the face (mouth, eyebrows, etc.) and analyzes movement, shape and texture composition of these regions to identify facial action units (*AU*s). Therefore, it is possible to track tiny movements of facial muscles in individuals' face and translate them into universal facial expressions like happiness, surprise, sadness, anger and others.

Very important is also the way, how the emotional representation is created. Afzal et al. [48] performed studies differences between human raters' judgments of emotional expressions and three automatically generated representations in video clips. First, the point-based display was created from the output of a commercial automatic face-tracker [49] on a black background. Second one, stick figure models, which connecting the automatically tracked feature-points using straight lines and sketching eyes using typical shape. The last one, 3D XFace Animation was a 3D animated facial expressions and displays created using XFace open source toolkit. Their experiences confirmed that

the human judgement is still the best one, with the highest recognition rates, followed by stick-figure models, point-light displays, and then XFace animations.

3.4.1 Existing Tools for Automatic Facial Emotion Recognition

Facial emotion recognition is one of the most important methods for nonverbal emotion detection. Several popular commercial packages offer specific facial image analysis tasks, including facial expression recognition, facial attribute analysis, and face tracking. We cite below few examples:

IntraFace (IF) [50], a publicly available software package offering:

- automated facial feature tracking,
- head pose estimation,
- facial attribute recognition,
- facial expression analysis from video,
- unsupervised synchrony detection to discover correlated facial behavior between two or more persons.

It also measure an audience reaction to a talk given or synchrony for smiling in videos of parent-infant interaction.

The Emotion API (Microsoft Azure)
Microsoft Azure proposes on-line API to recognize emotions in images and in videos [51]. The Emotion API permits input data directly as an image or as "bounding box" from Face API. In the output, it returns the confidence across a set of eight emotions: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. Emotion API is able to track how a person or a crowd responds to your content over time.

Emotion API provide the interface for C#, cURL, Java, JavaScript, PHP, Python, Ruby. It is also possible to implement these API in R [52].

Micro Expressions Training Tool
Ekman has created several training tools [53] to enhance understanding of emotions and relationships:

- Micro Expressions Training Tool (METT and intensive MEITT) to teach how to detect and interpret micro expressions,
- Subtle Expressions Training Tool (SETT) to learn how to see emotions as they develop. SETT provides foundational knowledge of how emotions emerge in just one region on the face.
- Micro Expressions Profile Training Tool (MEPT): you to identify micro expressions from different angles.

3.5 Emotion Detected by the Physiological and Motor Signals

Recently the physiological and motor data are accessible by IoT technology. People are interested in purchasing connected objects in order to monitor their healthy like heart rate, blood pressure, number of burned calories and analyze their movements. We can find a lot work for healthcare applications but at the same time, this technology can be

used to emotion detection. As an example, we can cite work done by Amira et al. [54] as a good example of using the emotion analyze for healthcare purpose. This work takes into consideration the emotional state of the peoples (stress, happiness, sadness, among) and analyze that using the appropriate AI tools to detect emotion, to categorize it and then analyze its impact on cardiovascular disease.

Physiological Signals

Automatic emotion recognition based on physiological signals is a key topic for many advanced applications (safe driving, security, mHealth, etc.). Main analyzed physiological signals useful for emotion detection and classification are:

- electromyogram (EMG) - recording of the electrical activity produced by skeletal muscles,
- galvanic skin response (GSR) - reflecting skin resistance, which varies with the state of sweat glands in the skin controlled by the sympathetic nervous system, where conductance is an indication of psychological or physiological state,
- respiratory volume (RV) - referring to the volume of air associated with different phases of the respiratory cycle,
- skin temperature (SKT) - referring to the fluctuations of normal human body temperature,
- blood volume pulse (BVP) - measures the heart rate,
- heart rate (HR),
- electrooculogram (EOG) - measuring the corneo-retinal standing potential between the front and the back of the human eye,
- photoplethysmography (PPG) - measuring blood volume pulse (BVP), which is the phasic change in blood volume with each heartbeat, etc.

The recognition of emotions based on physiological signals covers different aspects: emotional models, methods for generating emotions, common emotional data sets, characteristics used and choices of classifiers. The whole framework of emotion recognition based on physiological signals has recently been described by [55].

Decision-level weight fusion strategy for emotion recognition in multichannel physiological signals using MAHNOB-HCI database has been recently described [56]. Various classification tools may be used, including artificial neural networks (ANN), support vector machine (SVM), k-nearest neighbors (KNN), and many more. More advanced emotion representation models, including clustering of responses, are created for purposes of further studies and increased recognition accuracy.

Özerdem and Polat used EEG signal, discrete wavelet transform and machine learning techniques (multilayer perceptron neural network - MLPNN, and k-nearest neighbors - kNN algorithms) [57].

Research by Jang et al. used ECG, EDA, SKT and more sophistical machine learning (ML) algorithms [58]:

- linear discriminant analysis (LDA) - finds a linear combination of features that characterizes or separates two or more classes of objects or events,
- classification and regression trees (CART) - uses a decision tree as a predictive model to go from observations about an item represented in the branches, to conclusions about the item's target value represented in the leaves,

- self-organizing map (SOM) - a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples, called a map,
- Naïve Bayes algorithm - based on applying Bayes' theorem with strong (=naive) independence assumptions between the features,
- Support Vector Machine (SVM) - supervised learning models with associated learning algorithms analyzing data used for classification and regression analysis.

Kortelainen et al. showed results of combining physiological signals (heart rate variability parameters, respiration frequency) and facial expressions were studied [59].

Motor Signals

Body posture and movement is one of the most expressive modalities for humans. Researchers have recently started to exploit the possibilities for emotion recognition based on different motor signals. Zachartos et al. present well done analyze of emerging techniques and modalities related to automated emotion recognition based on body movement and describes application areas and notation systems and explains the importance of movement segmentation [60]. At the same time, this work outlines that this field still requires further studies. To recognize the emotion different postural, kinematic and geometrical feature are used. Tsui et al. use the keystroke typing patterns (usually on a standard keyboard) for automatic recognizing of emotional state [61].

Li et al. proposed another way to solve this problem by analyzing the human pattern of movement of the limbs (gait) recorded by Microsoft Kinect [62]. The gait pattern for every subject was extracted from 3-dimensional coordinates of 14 main body joints using Fourier transformation and Principal Component Analysis (PCA). To classify signals' features four ML algorithms were trained and evaluated:

- Naive Bayes (described above),
- Random Forests - an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees,
- LibSVM - an integrated software for support vector classification, regression, and distribution estimation,
- Sequential Minimal Optimization (SMO) - is an algorithm for solving the quadratic programming problem arising during the training of support vector machines.

They showed that human gait reflects the walker's emotional state.

Accuracy

Accuracy of the emotion recognition and classification based on physiological signals identification have improved significantly. Research by Goshvarpour et al. using HRV and PRV, and fusion rules (feature level, decision level) showed classification rates improved up to 92% (sensitivity: 95%, specificity: 83.33%) [63]. Previous studies have provided evidence for general recognition rate from 57,77% to 85,46% for different emotional states, that more higher set of analyzed signals and more recognized emotions provides better results [64–66]. High-observed recognition accuracy was also in research by Jang et al.: 84.7% [58].

Özerdem and Polat used EEG signal - resultant average overall accuracy was 77.14% (using MLPNN) and 72.92% (using kNN) [56]. The highest accuracy achieved by Li et al. was 80.5% [62]. Accuracy for combining heart rate variability, respiration frequency, and facial expressions was relatively low: 54.5% [59]. Results of the 2011 i2b2/VA/Cincinnati Medical Natural Language Processing Challenge showed that the best of 24 participants' teams achieved accuracy of 61.39% [67].

It should be noted that all algorithm must be personalized to each person in order to be reliable.

4 Related Project

4.1 openSMILE (Speech & Music Interpretation by Large-Space Extraction) [68]

Originally created in the scope of the European EU-FP7 research project SEMAINE (http://www.semaine-project.eu). OpenSMILE is a modular and flexible feature extractor for signal processing and machine learning applications. However, due to their high degree of abstraction, openSMILE components can also be used to analyze signals from other modalities, such as physiological signals, visual signals, and other physical sensors, given suitable input components.

4.2 openEAR (Emotion and Affect Recognition)

It consists of three major components: the core component is the SMILE (Speech and Music Interpretation by Large-Space Extraction) signal processing and feature extraction tool, which is capable generating >500 k features in real-time (Real-Time Factor (RTF) < 0.1), either from live audio input or from offline media [69]. The advantages of this solution are: open-source, multi-threaded, real-time emotion recognition framework providing an extensible, platform independent feature extractor implemented in C++, pre-trained models on six databases which are ready-to-use for on-line emotion and affect recognition, and supporting scripts for model building, evaluation, and visualization. This framework is compatible with related tool-kits, such as HTK and WEKA by supporting their data-formats.

4.3 ASC-Inclusion (Interactive Emotion Games for Children with Autism)

Understand and express emotions through facial expressions, vocal intonation and body gestures. This project aims to create an internet-based platform that will assist children with Autism Spectrum Condition (ASC) to improve their socio-emotional communication skills. The project will attend both the recognition and the expression of socio-emotional cues, aiming to provide an interactive-game where to give scores on the prototypically and on the naturalness of child's expressions. It will combine several state-of-the-art technologies in one comprehensive virtual world environment, combining voice, face and body gesture analysis, providing corrective feedback regarding the appropriateness of the child's expressions [70].

4.4 INTERSPEECH - Computational Paralinguistics ChallengE (ComParE)

In this Challenge, authors introduced four paralinguistic tasks, which are important for the realm of affective human-computer interaction, yet some of them go beyond the traditional tasks of emotion recognition. Thus, as a milestone, ComParE 2013 laid the foundation for a successful series of follow-up ComParEs to date, exploring more and more the paralinguistic facets of human speech in tomorrow's real-life information, communication and entertainment systems [71].

5 Conclusion

In this paper, we presented existing multimodal approaches and methods of emotion detection and analysis. Our study includes emotion analyze in text, in sound, in image/video and physiological signals. We showed that automated emotion analysis is possible and can be very useful for improving the exactness of the computer/machine reaction and making possible to anticipate the emotional state of the interlocutor more quickly.

Not only basic emotions can be analyzed, but also the cognitive assessment (interpretation of stimuli in the environment) and their physiological response. Such an approach may cause quicker development of more user-friendly systems and environments supporting everyday life and work.

Automatic emotion analysis requires advanced recognition and modeling, very often based on artificial intelligence systems. Presented approach may be successful, but the limitations of the current knowledge and experience still concern tools for automatic non-invasive emotion measurement and analysis.

We should be aware that among requirements on automatic emotion recognition key might constitute portability, non-intrusiveness, and low price. Novel intelligent systems may be friendlier, preventing the computers from acting inappropriately.

In this state of the arts, we have addressed, of course only some most important issues. No doubt, there is need for further effort of scientists and engineers toward more advanced AI-based automatic emotion recognition systems. Further research may constitute an important part of future technological, clinical, and scientific progress.

Acknowledgement (COST + PHC Polonium). This state of the art was published in cooperation with the COST Action IC1406 High-Performance Modelling and Simulation for Big Data Applications (cHiPSet), supported by COST (European Cooperation in Science and Technology).

This article is based upon the work done under the project PHC POLONIUM 2018 (PROJECT N° 40557VL), realized between the AlliansTIC Research Laboratory of Efrei Paris Engineering School (France) and the Institute of Mechanics and Applied Computer Science, Kazimierz Wielki University in Bydgoszcz (Poland).

References

1. Mehrabian, A., Ferris, S.R.: Inference of attitudes from nonverbal communication in two channels. J. Consult. Psychol. **31**(3), 248 (1967)
2. Mood Ring Monitors Your State of Mind, Chicago Tribune, 8 October 1975, at C1: Ring Buyers Warm Up to Quartz Jewelry That Is Said to Reflect Their Emotions. The Wall Street Journal, 14 October 1975, at p. 16; and "A Ring Around the Mood Market", The Washington Post, 24 November 1975, at B9
3. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
4. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. **23**(10), 1175–1191 (2001)
5. Hernandez, J., et al.: AutoEmotive: bringing empathy to the driving experience to manage stress. In: DIS 2014, 21–25 June 2014, Vancouver, BC, Canada. ACM (2014). http://dx.doi.org/10.1145/2598784.2602780. 978-1-4503-2903-3/14/06
6. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell. Syst. **31**(6), 82–88 (2016). https://doi.org/10.1109/mis.2016.94
7. Wöllmer, M., et al.: YouTube movie reviews: sentiment analysis in an audio-visual context. IEEE Intell. Syst. **28**(3), 46–53 (2013)
8. Perez-Rosas, V., Mihalcea, R., Morency, L.P.: Utterance-level multimodal sentiment analysis. In: ACL, vol. 1, pp. 973–982 (2013)
9. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis, arXiv:1707.07250. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 7–11 September 2017, Copenhagen, Denmark, pp. 1103–1114. Association for Computational Linguistics
10. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P.: Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 873–883 (2017)
11. Poria, S., Cambria, E., Howard, N., Huang, G.B., Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing **174**(Part A), 50–59 (2016). https://doi.org/10.1016/j.neucom.2015.01.095. ISSN 0925-2312
12. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
13. Pang, B., Lee, L.: Opinion mining and sentiment analysis. J. Found. Trends Inf. Retrieval **2** (1–2), 1–135 (2008)
14. Dziczkowski, G., Wegrzyn-Wolska, K.: RRSS - rating reviews support system purpose built for movies recommendation. In: Wegrzyn-Wolska, K.M., Szczepaniak, P.S. (eds.) Advances in Intelligent Web Mastering. Advances in Soft Computing, vol. 43, pp. 87–93. Springer, Berlin (2007). https://doi.org/10.1007/978-3-540-72575-6_14
15. Dziczkowski, G., Węgrzyn-Wolska, K.: An autonomous system designed for automatic detection and rating of film. Extraction and linguistic analysis of sentiments. In: Proceedings of WIC, Sydney (2008)
16. Dziczkowski, G., Węgrzyn-Wolska, K.: Tool of the intelligence economic: recognition function of reviews critics. In: ICSOFT 2008 Proceedings. INSTICC Press (2008)
17. Kepios: Digital in 2018, essential insights into internet, social media, mobile, and ecommerce use around the world, April 2018. https://kepios.com/data/
18. Ghiassi, M., Skinner, J., Zimbra, D.: Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Expert Syst. Appl. **40**(16), 6266–6282 (2013)

19. Zhou, X., Tao, X., Yong, J., Yang, Z.: Sentiment analysis on tweets for social events. In: Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2013, 27–29 June 2013, pp. 557–562 (2013)
20. Salathé, M., Vu, D.Q., Khandelwal, S., Hunter, D.R.: The dynamics of health behavior sentiments on a large online social network. EPJ Data Sci. **2**, 4 (2013). https://doi.org/10.1140/epjds16
21. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in Twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 19–23 July 2010, pp. 841–842. http://doi.acm.org/10.1145/1835449.1835643
22. Seabrook, E.M., Kern, M.L., Fulcher, B.D., Rickard, N.S.: Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. J. Med. Internet Res. **20**(5), e168 (2018). https://doi.org/10.2196/jmir.9267
23. Wang, W., Hernandez, I., Newman, D.A., He, J., Bian, J.: Twitter analysis: studying US weekly trends in work stress and emotion. Appl. Psychol. **65**(2), 355–378 (2016)
24. Reece, A.G., Reagan, A.J., Lix, K.L., Dodds, P.S., Danforth, C.M., Langer, E.J.: Forecasting the onset and course of mental illness with Twitter data (Unpublished manuscript). https://arxiv.org/pdf/1608.07740.pdf
25. Park, J., Lee, D.S., Shablack, H., et al.: When perceptions defy reality: the relationships between depression and actual and perceived Facebook social support. J. Affect. Disord. **200**, 37–44 (2016)
26. Burke, M., Develin, M.: Once more with feeling: supportive responses to social sharing on Facebook. In: Proceedings of the ACM 2016 Conference on Computer Supported Cooperative Work, pp. 1462–1474 (2016)
27. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. J. CS224N Proj. Rep., Stanford **1**, 12 (2009)
28. Liu, K.L., Li, W.J., Guo, M.: Emoticon smoothed language models for Twitter sentiment analysis. In: AAAI (2012)
29. Węgrzyn-Wolska, K., Bougueroua, L., Yu, H., Zhong, J.: Explore the effects of emoticons on Twitter sentiment analysis. In: Proceedings of Third International Conference on Computer Science & Engineering (CSEN 2016), 27–28 August 2016, Dubai, UAE
30. Bitouk, D., Verma, R., Nenkova, A.: Class-level spectral features for emotion recognition. Speech Commun. **52**(7–8), 613–625 (2010)
31. Busso, C., et al.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Sixth International Conference on Multimodal Interfaces, ICMI 2004, October 2004, State College, PA, pp. 205–211. ACM Press (2004)
32. Dellaert, F., Polzin, T., Waibel, A.: Recognizing emotion in speech. In: International Conference on Spoken Language (ICSLP 1996), October 1996, Philadelphia, PA, USA, vol. 3, pp. 1970–1973 (1996)
33. Lee, C.M., et al.: Emotion recognition based on phoneme classes. In: 8th International Conference on Spoken Language Processing (ICSLP 2004), October 2004, Jeju Island, Korea, pp. 889–892 (2004)
34. Deng, J., Xu, X., Zhang, Z., Frühholz, S., Grandjean, D., Schuller, B.: Fisher kernels on phase-based features for speech emotion recognition. In: Jokinen, K., Wilcock, G. (eds.) Dialogues with Social Robots. LNEE, vol. 427, pp. 195–203. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-2585-3_15
35. Steidl, S.: Automatic classification of emotion-related user states in spontaneous children's speech. Ph.D. thesis, Erlangen (2009)
36. Lugovic, S., Horvat, M., Dunder, I.: Techniques and applications of emotion recognition in speech. In: MIPRO 2016/CIS (2016)

37. Kukolja, D., Popović, S., Horvat, M., Kovač, B., Ćosić, K.: Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications. Int. J. Hum.-Comput. Stud. **72**(10), 717–727 (2014)
38. Davletcharova, A., Sugathan, S., Abraham, B., James, A.P.: Detection and analysis of emotion from speech signals. Procedia Comput. Sci. **58**, 91–96 (2015)
39. Tyburek, K., Prokopowicz, P., Kotlarz, P.: Fuzzy system for the classification of sounds of birds based on the audio descriptors. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2014. LNCS (LNAI), vol. 8468, pp. 700–709. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07176-3_61
40. Tyburek, K., Prokopowicz, P., Kotlarz, P., Michal, R.: Comparison of the efficiency of time and frequency descriptors based on different classification conceptions. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2015. LNCS (LNAI), vol. 9119, pp. 491–502. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19324-3_44
41. Chaspari, T., Soldatos, C., Maragos, P.: The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences. World J. Biol. Psychiatry **16**(5), 312–322 (2015)
42. Arruti, A., Cearreta, I., Alvarez, A., Lazkano, E., Sierra, B.: Feature selection for speech emotion recognition in Spanish and Basque: on the use of machine learning to improve human-computer interaction. PLoS ONE **9**(10), e108975 (2014)
43. Ekman, P.: Facial expression and emotion. Am. Psychol. **48**, 384–392 (1993)
44. Jack, R.E., Schyns, P.G.: The human face as a dynamic tool for social communication. Curr. Biol. Rev. **25**(14), R621–R634 (2015). https://doi.org/10.1016/j.cub.2015.05.052
45. Ekman, P., Friesen, W., Hager, J.: Facial action coding system: Research Nexus. Network Research Information, Salt Lake City (2002)
46. Hjorztsjö, C.H.: Man's face and mimic language (1969). https://books.google.com/books/about/Man_s_Face_and_Mimic_Laguage.html?id=BakQAQAAIAAJ
47. Ekman, P., Huang, T.S., Sejnowski, T.J., et al.: Final report to NSF of the planning workshop on facial expression understanding, vol. 378. Human Interaction Laboratory, University of California, San Francisco (1993)
48. Afzal, S., Sezgin, T.M., Gao, Y., Robinson, P.: Perception of emotional expressions in different representations using facial feature points. IEEE (2009). 978-1-4244-4799
49. http://www.nevenvision.com. Licensed from Google Inc.
50. De la Torre, F., Chu, W.S., Xiong, X., Vicente, F., Ding, X., Cohn, J.: IntraFace. In: IEEE International Conference on Automatic Face and Gesture Recognition Workshops (2015). https://doi.org/10.1109/fg.2015.7163082
51. https://azure.microsoft.com/en-us/services/cognitive-services/emotion/
52. http://thinktostart.com/analyze-face-emotions-r/
53. https://www.paulekman.com/micro-expressions-training-tools/
54. Amira, T., Dan, I., Az-Eddine, B., et al.: Monitoring chronic disease at home using connected devices. In: 2018 13th Annual Conference on System of Systems Engineering (SoSE), pp. 400–407. IEEE (2018)
55. Shu, L., et al.: A review of emotion recognition using physiological signals. Sensors (Basel) **18**(7), 2074 (2018)
56. Wei, W., Jia, Q., Feng, Y., Chen, G.: Emotion recognition based on weighted fusion strategy of multichannel physiological signals. Comput. Intell. Neurosci. **2018**, 9 (2018). 5296523
57. Özerdem, M.S., Polat, H.: Emotion recognition based on EEG features in movie clips with channel selection. Brain Inform. **4**(4), 241–252 (2017)
58. Jang, E.H., Park, B.J., Park, M.S., Kim, S.H., Sohn, J.H.: Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. J. Physiol. Anthropol. **34**, 25 (2015)

59. Kortelainen, J., Tiinanen, S., Huang, X., Li, X., Laukka, S., Pietikäinen, M., Seppänen, T.: Multimodal emotion recognition by combining physiological signals and facial expressions: a preliminary study. In: Conference Proceeding of the IEEE Engineering in Medicine and Biology Society, vol. 2012, pp. 5238–5241 (2012)
60. Zacharatos, H., Gatzoulis, C., Chrysanthou, Y.L.: Automatic emotion recognition based on body movement analysis: a survey. IEEE Comput. Graph Appl. **34**(6), 35–45 (2014)
61. Tsui, W.H., Lee, P., Hsiao, T.C.: The effect of emotion on keystroke: an experimental study using facial feedback hypothesis. In: Conference Proceedings of the IEEE Engineering in Medicine and Biology Society, pp. 2870–2873 (2013)
62. Li, S., Cui, L., Zhu, C., Li, B., Zhao, N., Zhu, T.: Emotion recognition using Kinect motion capture data of human gaits. PeerJ **4**, e2364 (2016)
63. Goshvarpour, A., Abbasi, A.: Goshvarpour, A: Fusion of heart rate variability and pulse rate variability for emotion recognition using lagged poincare plots. Australas. Phys. Eng. Sci. Med. **40**(3), 617–629 (2017)
64. Khezri, M., Firoozabadi, M., Sharafat, A.R.: Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals. Comput. Methods Programs Biomed. **122**(2), 149–164 (2015)
65. Gouizi, K., Bereksi Reguig, F., Maaoui, C.: Emotion recognition from physiological signals. J. Med. Eng. Technol. **35**(6–7), 300–307 (2011)
66. Verma, G.K., Tiwary, U.S.: Multimodal fusion framework: a multiresolution approach for emotion classification and recognition from physiological signals. Neuroimage **102**(Part 1), 162–172 (2014)
67. Yang, H., Willis, A., de Roeck, A., Nuseibeh, B.: A hybrid model for automatic emotion recognition in suicide notes. Biomed. Inform. Insights **5**(Suppl. 1), 17–30 (2012)
68. Eyben, F., Weninger, F., Wöllmer, M., Shuller, B.: Open-Source Media Interpretation by Large Feature-Space Extraction, November 2016. openSMILE by audFERING
69. Eyben, F., Wöllmer, M., Shuller, B.: openEAR - introducing the munich open-source emotion and affect recognition toolkit. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. https://doi.org/10.1109/acii.2009.5349350
70. O'Reilly, H., et al.: The EU-emotion stimulus set: a validation study. Behav. Res. **48**, 567–576 (2016). https://doi.org/10.3758/s13428-015-0601-4. Psychonomic Society, Inc. 2015
71. Schuller, B., et al.: Affective and behavioural computing: lessons learnt from the first computational paralinguistics challenge. Comput. Speech Lang. **53**, 156–180 (2019). Elsevier, ScienceDirect