HAL
open science

Fast 3D keypoints detector and descriptor for view-based 3D objects recognition

Ayet Shaiek, Fabien Moutarde

To cite this version:

Ayet Shaiek, Fabien Moutarde. Fast 3D keypoints detector and descriptor for view-based 3D objects recognition. International Workshop on Depth Image Analysis (WDIA'2012) of 21st International Conference on Pattern Recognition (ICPR'2012), Nov 2012, Tsukuba, Japan. hal-00766679

HAL Id: hal-00766679

https://minesparis-psl.hal.science/hal-00766679v1

Submitted on 18 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Fast 3D keypoints detector and descriptor for view-based 3D objects recognition

Ayet Shaiek[1], and Fabien Moutarde[1]

[1]Robotics laboratory (CAOR) Mines ParisTech 60 Bd St Michel, F-75006 Paris, France

Abstract. In this paper, we propose a new 3D object recognition method that employs a set of 3D keypoints extracted from point cloud representation of 3D views. The method makes use of the 2D organization of range data produced by 3D sensor. Our novel 3D interest points approach relies on surface type classification and combines the Shape Index (SI) - curvedness(C) map with the Gaussian (H) - Mean (K) map. For each extracted keypoint, a local description using the point and its neighbors is computed by joining the Shape Index histogram and the normalized histogram of angles between normals. This new proposed descriptor IndSHOT stems from the descriptor CSHOT (Color Signature of Histograms of OrienTations) which is based on the definition of a local, robust and invariant Reference Frame RF. This surface patch descriptor is used to find the correspondences between query-model view pairs in effective and robust way. Experimental results on Kinect based datasets are presented to validate the proposed approach in view based 3D object recognition.

Keywords: Depth Image, 3D Keypoints detector, Mean Curvature, Gaussian Curvature, Shape Index, HK Map, SC Map, SHOT Descriptor, IndSHOT.

1 Introduction

There has been strong research interest in 3D object recognition over the last decade, due to the promising reliability of the new 3D acquisition techniques. 3D recognition, however, conveys several issues related to the amount of information, class variability, partial information, as well as scales and viewpoints differences are encountered. As previous works in the 2D case have shown, local methods perform better than global features to partially overcome those problems. Global features need the complete, isolated shape for their extraction. Examples of global 3D features are volumetric part-based descriptions [1]. These methods are less successful when dealing with partial shape and intra-class variations while remaining partially robust to noise, clutter and inter-class variations. The field of 2D Point-of-interest (POI) feature has been the source of inspiration for the 3D interest-points detectors. For example, the Harris detector has been extended to three dimensions, first in [2] with two spatial dimen-

sions plus the time dimension, then in [3] which discusses variants of the Harris measure and recently in [4] where a 3D-SURF adaptation is proposed. The 3D shape of a given object can be described by a set of local features extracted from patches around salient interest points. Regarding efficient 3D descriptors, the SHOT descriptor [5] achieves both state-of-the-art robustness, and descriptiveness. Results demonstrate the higher descriptiveness embedded in SHOT with respect to Spin Images [6] Exponential Mapping (EM) and Point Signatures (PS). Given the local RF, an isotropic spherical grid is defined to encode spatially well localized information. For each sector of the grid a histogram of normals is defined and the overall descriptor SHOT results from the juxtaposition of these histograms.

Our proposed new method aims to detect salient keypoints that are repeatable under moderate viewpoint variations. We propose to use a measure of curvature in the line of Chen and Bhanu's work [7] and construct a patch labeling to classify different surface shapes [7, 8] using both mean-gaussian curvatures (HK) and shape index-curvedness (SC) couples. Thus, we select keypoints according to their local surface saliency. Furthermore, we suggest a novel descriptor, dubbed IndSHOT, that emphasizes the shape description by merging the SHOT descriptor with the Shape Index histogram. The complete recognition system with detection, description and matching phases is introduced in §2. The proposed method is then evaluated in §3.

2 Methodology

2.1. Resampling of the 3D Points Cloud

As we address a recognition scenario wherein only 2.5 views are matched, we deal with some views of the models from specific viewpoints. In the work presented here, we exploit the lattice structure provided by the range image. First, we search the coordinates of the maximum and minimum points at x-axis and y-axis in the sample, and build a bounding box based on the two limit points. Using the (i,j) coordinates of each point in this box, we smooth the initial 3D point cloud by resampling down to 1/span of its original point density in order to avoid noise perturbation. Then, we generate a mesh using the new vertices corresponding to the average of points belonging to a rectangular region with a span in the x and y direction. The x and y spans are proportional to the density of points and to a fraction r_1 of the bounding box dimensions, so as to make our method robust to different spatial samplings and to scaling. In our approach, neighbour points are given by a spherical region around the point, with a support radius $R = r_2$ x mesh-resolution. In practice, we adjust a local polynomial surface to the selected neighborhood. CGAL library is used for curvature computation. An advantage of subdividing the point cloud in local regions is to avoid mutual impact between them.

2.2. Keypoint Detectors

The aim of this step is to pick out a repeatable and salient set of 3D points. Principal curvatures correspond to the eigenvalues of the Hessian matrix and are invariant under rotation. Hence, we propose to use local curvatures which can be calculated either directly from first and second derivatives, or indirectly as the rate of change of normal orientations in a local context region. The usual pair of Gaussian curvature K and mean curvature H only provides a poor representation, since the values are strongly correlated. Instead, we use them in composed form with curvature based quantities. In

the following, we, first, introduce state-of-the-art detector methods based on shape index, HK and SC classification; then we present the principle of our new detector.

Shape Index. This detector type was proposed in [7], and uses the shape index (SIp) for feature point extraction. It is a quantitative measure of the surface shape at a point p, and is defined by (eq. 1),

$$SI_p = \frac{1}{2} - \frac{1}{\pi} \times \mathrm{arctg}\left(\frac{k_p^1 + k_p^2}{k_p^1 - k_p^2}\right) \quad (1)$$

where k^1_p and k^2_p are maximum and minimum principal curvatures, respectively. With this definition, all shapes are mapped into the interval [0, 1] where every distinct surface shape corresponds to a unique value of SI (except for planar surfaces, which will be mapped to the value 0.5, together with saddle shapes). Larger shape index values represent convex surfaces and smaller shape index values represent concave surfaces. The main advantage of this measure is the invariance to orientation and scale. A point is marked as a feature point if its shape index SI_p satisfies (2) within point neighbors,

$$\begin{cases} SI_p = max(SI_k)\,; k \in neighbors \;\; and \;\; SI_p \geq \; (1+\alpha) \times \mu \\ \qquad\qquad or \\ SI_p = min(SI_k)\,; k \in neighbors \;\; and \;\; SI_p \leq \; (1-\beta) \times \mu \end{cases} \quad (2)$$

where μ is the mean of shape index over the SI point neighbors values and $0 \leq \alpha$, $\beta \leq 1$. In above expression (eq. 2), α and β parameters control the selection of feature points. We denote this detector by « SID ».

HK and SC Classification. The idea here is to build shape classification space using the pair mean-Gaussian curvatures (HK) or the pair shape index-curvedness (SC). Typically, for HK classification, we use the type function Tp used in LSP descriptor [7] that associates to each couple of H and K values a unique type value (eq. 3),

$$\mathrm{T_p} = 1 + 3\left(1 + \mathrm{sgn}_{\varepsilon_H}(H)\right) + \left(1 - \mathrm{sgn}_{\varepsilon_K}(K)\right);\; \mathrm{sgn}_{\varepsilon_X}(X) \begin{cases} +1 & if\; X > \varepsilon_X, \\ 0 & if\; |X| \leq \varepsilon_X, \\ -1 & if\; X < \varepsilon_X \end{cases} \quad (3)$$

where ε_H and ε_K are two thresholds over the H and K. Nine region types are defined.

In the shape index-curvedness (SC) space, S defines the shape type and C defines the degree of curvature and is the square-root of the deviation from flatness. Similarly to HK representation, the continuous graduation of S subdivides surface shapes into 9 types. Planar surfaces are classified using the C value. We define a type function S_p (eq. 4) that associates a unique type value to each couple of SI and C values (i.e values between 0.8125 and 0.9375 correspond to dome and $S_p = 7$),

$$\begin{cases} S_p = 0 \text{ if } C \leq \varepsilon_C \\ \text{else} \\ S_p \in [1,8\,]\;;\; SI \in [\![0,1]\!]\,. \end{cases} \quad (4)$$

For both classifications, salient regions are selected as those of one of the 5 following types: dome, trough, spherical, saddle rut and saddle ridge regions. More details are given in [9, 10].

Combination of Criteria. Theoretically, the two classifications HK and SC should provide the same result; therefore we suggest combining the two criteria to increase reliability. In fact, our result will be validated with two measures of keypoints detection. After labeling points with a pair of value (T_p, S_p), points with salient type pair are selected, in other words, if the two labels correspond to the same of the 5 salient region types previously mentioned. We note this detector « SC_HK ». Then, points with the same pair value are grouped using the connected- component labeling. Connectivity is carried out by checking the 8-connectivity of each point. Finally, the centers of the connected component are selected as keypoints. We also propose further combination by ranking the selected keypoints according to their curvedness value. The point with the maximum value of curvedness over the selected keypoints is chosen to represent each connected component. We call the detector combining the two criteria « SC_HK_connex ».

2.1 Keypoint Descriptors

After keypoints detection step, a 3D descriptor is computed around each selected interest point. In the case of range data, the dominant orientation at a point is the direction of the surface normal at that point. Histogram-based methods are typically based on the feature point normals. For example, Local Surface Patches [7] computes histograms of normals and shape indexes of the points belonging to the keypoint support. The recently proposed SHOT descriptor achieves computational efficiency, descriptive power and robustness by defining 3D repeatable local Reference Frame (RF). We briefly summarize here the structure of the SHOT descriptor. The reader is referred to [5] for details on the descriptor. The introduction of geometric information concerning the location of the points within the support is performed by first calculating a set of local histograms of normals over the 3D volumes defined by a 3D grid superimposed on the support and then grouping together all local histograms to form the final descriptor. The normal estimation is based on the Eigenvalue Decomposition of a novel scatter matrix defined by a weighted linear combination of neighbour point distances to the feature point, lying within the spherical support. The eigenvectors of this matrix define repeatable, orthogonal directions in presence of noise and clutter. Furthermore, the CSHOT descriptor [11] is proposed as an amelioration of the SHOT descriptor and makes profits from the 3D data enriched with texture. The process of combination succeeds to form more robust and descriptive signature.

Inspired from theses state-of-the-art descriptors, we compute the histograms of shape index values and of angle values between the reference surface normals at the feature point and the neighbour's ones and join the two histograms similarly to the design of CSHOT descriptor. First of all, we accumulate point counts into bins according to a cosine function of the angle between the normal at each point within the corresponding part of the grid and the normal at the feature point. For each of the local histograms, a coarser binning is created for directions close to the reference normal direction and a finer one for orthogonal directions. In this way, small differences in orthogonal directions to the normal, which are the most informative ones, cause a point to be accumulated in different bins. Secondly, shape index values of the feature point and those of its neighbours relying in the spherical support are grouped into bins. Finally, we merge the shape index values and the cosine values into one descriptor that we call IndSHOT. We perform the same process as in the CSHOT to juxtapose the two histograms, where index shape histogram replaces the color histogram. In

addition, the mean and standard deviation of shape index of the neighbors around the feature point are computed. The final descriptors, composed of (model ID, index shape + cosines histograms, surface type, the 3D coordinates of keypoint, mean and standard deviation of shape index), are saved to be used in the matching process.

2.2 Matching and Recognition

We are validating the proposed detector and descriptor using a view matching approach. Here, we focus on solving the surface matching problem based on local features, by point-to-point correspondences obtained by matching local invariant descriptors of feature points. Given a test object, we compute a measure of similarity between descriptors extracted on the test view and those of the models in database. The information (model ID, histogram, surface type, the centroïd, mean and standard deviation of SI) are used for matching process. Hence, for each histogram from test view, we find the best matching histogram from database view using the Euclidian distance. To speed up the comparison process, we use a KD-tree structure. Two keypoints are matched according to their histogram distance and their types of surface. For a test object, a set of nearest neighbors is returned after histogram matching. In the case of multiple correspondences, the potential corresponding pairs are filtered based on the geometric constraint: Euclidean distance between features coordinates of the two matched surface patches. The closest couple of features in term of coordinates distance is the more likely to form a consistent correspondence. A system of incremental votes for each class gives the final matched class.

3 Experimental results

3.1. Data and Parameters

We performed our experiments on two real range data sets. The first one is the public RGB-D Object Dataset [12] (figure 1). There are 51 common household object categories. In our experimentation, we use 46 objects with 25 views per object for only one object per category, which constitute a dataset of 1150 views. The list of the following objects are labelled from 1 to 46 respectively: apple_1, ball_1, banana_1, bell_peper_1, binder_1, calculator_1, camera_1, cap_1, cell_phone_1, cereal_box_3, coffee_mug_1, comb_1, flashlight_1, food_bag_1, food_box_1, food_can_1, food_cup_1, garlic_1, greens_1, hand_towel_1, instant_noodles_1, keyboard_1, Kleenex_1, lemon_1, lightbulb_1, lime_1, marker_1, mushroom_1, notebook_1, onion_1, orange_1, peach_1, pear_1, pitcher_1, plate_1, potato_1, rubber_eraser_1, scissors_1, shampoo_1, soda_can_1, sponge_1, stapler_1, tomato_1, toothbrush_1 and watter_bottle_1. The second data set is our own dataset (Lab-dataset) captured with the Kinect sensor and composed of 20 objects (Ex. prism, ball, fan, trash can, etc) with 3 to 10 different angle views per object (figure 2). The numbers of feature points detected from these range images vary from 4 to 250, depending on the viewpoint and the complexity of input shape. The parameters of our approach are: $r_1 = 2$, $r_2 = 4$, $\boldsymbol{\alpha} = 0.05$, $\beta = 0.05$, $\varepsilon_H = 0.009$, $\varepsilon_K = 0.0001$, $\varepsilon_C = 0.01$.

Fig. 1. *Examples of objects from the RGB-D Object Dataset [12]*

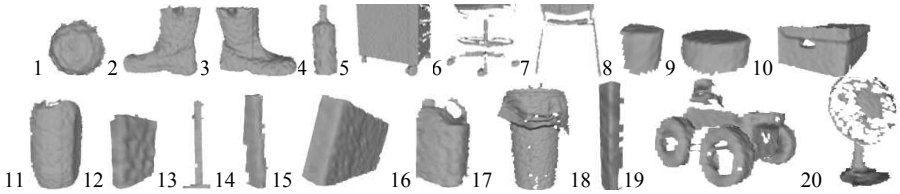

Fig. 2. *The 20 objects of the lab-Dataset*

3.2. Keypoint Stability

To evaluate detector performance, we illustrate a visual comparison of keypoint positions detected with SC_HK , SC_HK_connex, and SID detectors as shown on figure 3. It reveals that the final selected points are quite well localized. The combining process allows a better feature point filtering than SC or HK alone, as false detected points in both are eliminated, and points with correct surface type remain. Figure 4 illustrates the relative stability of keypoint's positions detected with SC_HK_connex detector when varying viewpoints for the same object. Clearly, we recover almost same keypoint positions in the different views. For a quantitative analysis showing the superior repeatability of our keypoints, we refer the reader to our previous publication [13].

Fig. 3. *Detected keypoints on trash can, fan and storage cupboard models with: SID in first column, SC_HK in second column and SC_HK_connex in third column.*

Fig. 4. *Detected keypoint on fan model with SC_HK_connex, in view angle variation*

3.3. Matching Result

The test protocol for object recognition from different angle views is the following: for the RGBD dataset, we select one test view from the N total number of views in the dataset, and the N-1 views are used as the training set; this process is repeated for the N views of the whole database. For the lab-dataset, we select one to four random views per object as the query and use the remaining views for training. We carry out three experiments using the three descriptors SHOT, CSHOT and IndSHOT. The same evaluation is done for the two detectors SID and SC_HK_connex. The Overall recognition rates are given in table 1 for respectively our lab dataset and the RGBD dataset. In figure 5, the cross recognition rates between models are displayed in the

confusion matrix. Gray level determines the rate of the recognition. Black is for high and white is for low recognition rate. The overall recognition rate is quite promising for our SC_HK_connex method in comparison to the SID results, with 91.12% on the RGBD dataset. This rate is achieved using the new proposed descriptor IndSHOT, which suggests that it is more descriptive than the CSHOT and SHOT versions. The recognition rate in the Lab dataset is about 82%. The reason behind this lower result is the high similarity between object shapes included in this dataset (two boots objects, parallelepipedic shapes, cylindrical shapes, etc).

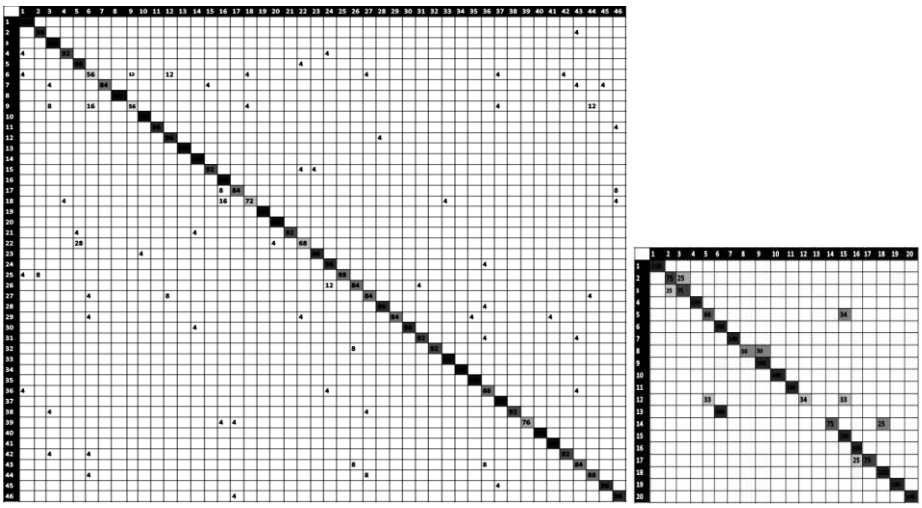

Fig. 5. *Confusion matrix for the result of SC_HK_connex method on RGB-D object dataset (on left) and on the Lab-dataset (on right).*

Table1: *Recognition rates for our Lab-dataset (on left) and RGB-D object dataset (on right)*

	IndSHOT	**SHOT**	**CSHOT**
SC_HK	**82.5%**	62,5%	67.5%

	IndSHOT	**SHOT**	**CSHOT**
SID	89.06%	70,75%	77.77%
SC_HK	**91.12%**	75,28%	82.14%

The conjunction of the SC_HK_connex detector with the IndSHOT descriptor seems to provide more pertinent description of the local surface typology. It should also be noted that the overall computation time for our recognition process (detection+ description+ matching features) is quite low (~0.7s), which is a great advantage when dealing with real time application.

4. Conclusions and Perspectives

We have presented two main complementary contributions: 1/ an original 3D keypoint detector, SC_HK_connex, based on the idea of combining criteria; 2/ a new 3D keypoint descriptor, IndSHOT, based solely on shape characteristics.

The proposed detector combines SC (shape curvedness) and HK criteria with the principle of connected components. It was already shown in our previous work that the selected 3D keypoints are more repeatable than for alternative detectors, and this is confirmed here by the good inter-view matching reached in our experiments. The proposed IndSHOT descriptor encodes the occurrence frequency of shape index values vs. the cosine of the angle between the normal of reference feature point and that of its neighbours. It seems to be significantly more descriptive than original SHOT and CSHOT from which we have crafted it.

Finally, our new combination of SC_HK_connex detector + IndSHOT descriptor is evaluated in challenging 3D object recognition scenarios characterized by the presence of viewpoint variations and a few number of views on real-world depth data. The outcome is very promising results, with 91% correct recognition on 46 objects from a public dataset, and 82% on our own lab dataset containing 20 “everyday” objects, some of which are rather similar one to another.

For the moment, measures of curvatures in our process are calculated at a constant scale level, so the feature’s scale is still ambiguous. To overcome this fact, we plan, as a future work, to search for features at different scale levels.

References

[1] Medioni, G.G. and François, A.R.J. "3-D structures for generic object recognition," Computer Vision and Image Analysis, 1, 1030 (2000).

[2] Paul S., Saad A. and Mubarak S., "A 3-dimensional SIFT descriptor and its application to action recognition", Proceedings of the 15th International Conference on Multimedia, 357–360 (2007).

[3] Fredrik V., Klas N., Mikael K. "Point-of-Interest Detection for Range Data", ICPR IEEE, 1-4 (2008).

[4] Jan K., Mukta P., Geert W., Radu T., Luc van G., "Hough Transform and 3D SURF for robust three dimensional classification," Proceedings of the European Conference on Computer Vision, (2010).

[5] Samuele S., Federico T., and Luigi Di S. "Unique Signatures of Histograms for Local Surface Description," in Proc. ECCV, (2010).

[6] Johnson, A.E., Hebert, M., "Using spin images for efficient object recognition in cluttered 3d scenes," IEEE PAMI 21, 433-449 (1999).

[7] Chen, H. and Bhanu, B. "3D free-form object recognition in range images using local surface patches," Pattern Recognition Letters, 28(10), 1252-126 (2007).

[8] Erdem A., Omer E., Ilkay U. "Scale-space approach for the comparison of HK and SC curvature descriptions as applied to object recognition". ICIP, 413-416 (2009).

[9] H. Cantzler, R. B. Fisher, "Comparison of HK and SC curvature description methods" In Conference on 3D Digital Imaging and Modeling, 285-291 (2001).

[10] J. Koenderink and A. J. Doorn. "Surface shape and curvature scale ", Image Vis. Comput., vol. 10, no. 8, pp. 557–565, (1992).

[11] Federico T., Samuele S., and Luigi Di S. "A combined texture-shape descriptor for enhanced 3D feature matching"; IEEE International Conference on Image Processing (ICIP), Brussels, Belgium, September 11-14 (2011).

[12] http://www.cs.washington.edu/rgbd-dataset/

[13] Ayet S. and Fabien M., "Détecteurs de points d’intérêt 3D basés sur la courbure" In COmpression et REprésentation des Signaux Audiovisuels (CORESA), 2012.