3D keypoint detectors and descriptors for 3D objects recognition with TOF camera

Ayet Shaiek [*a], Fabien Moutarde [†a]
[a]Robotics laboratory (CAOR) Mines ParisTech 60 Bd St Michel, F-75006 Paris, France;

ABSTRACT

The goal of this work is to evaluate 3D keypoints detectors and descriptors, which could be used for quasi real time 3D object recognition. The work presented has three main objectives: extracting descriptors from real depth images, obtaining an accurate degree of invariance and robustness to scale and viewpoints, and maintaining the computation time as low as possible. Using a 3D time-of-flight (ToF) depth camera, we record a sequence for several objects at 3 different distances and from 5 viewpoints. 3D salient points are then extracted using 2 different curvatures-based detectors. For each point, two local surface descriptors are computed by combining the shape index histogram and the normalized histogram of angles between the normal of reference feature point and the normals of its neighbours. A comparison of the two detectors and descriptors was conducted on 4 different objects. Experimentations show that both detectors and descriptors are rather invariant to variations of scale and viewpoint. We also find that the new 3D keypoints detector proposed by us is more stable than a previously proposed Shape Index based detector.

Keywords: 3D Keypoint Detector, 3D Keypoint Desciptor, Shape Index, Histogram Of Normals, Point Clouds, Time-Of-Flight Camera, Depth Image, 3D Objects Recognition.

1. INTRODUCTION

3D object recognition, an important research field, has been successfully studied in the case of a single viewpoint. Robustness to pose and viewpoint variations remains a challenging problem for objects recognition applications. Meanwhile, using new devices, such as time-off-light (TOF) 3D cameras, may be a step forward to provide robust geometric information about objects.

In this context, two components of an object recognition system are necessary: descriptions extraction phase where an interpretation of the image data is given, and the matching phase which consists of assigning an identity to the extracted descriptions.

The existing approaches for solving this issue can be classified in two ways:

- A group of 3D methods which suggest the use of the entire 3D model and base the recognition on the comparison of estimated model with reference models [1], and a group of 2D/3D approaches that project the 3D model into different 2D images [2, 3] and compute 2D features. A survey of 3D and multi-modal 3D+ 2D approaches has been done in [4].
- A class of global methods , like the work of [5] which suggest to use volumetric part-based descriptions, and a class of local ones which describe local regions, as for example in [6] , and in [7] where Viola and Jones propose a set of "rectangle" features.

Local representations using keypoint have been proven to perform well in 2D recognition [8]. Therefore, many recent researches have investigated in finding 3D detector and descriptor (eg: SIFT 2.5 [9]).

The proposed approach is in the line of Chen and Bhanu's work [10] who presented a local surface descriptor for 3D object recognition. In the work presented here, we use a depth camera "ZCam" (Figure 1) to capture four real objects (Figure 2)

[*] Ayet.Shaiek@mines-paristech.fr
[†] Fabien.Moutarde@mines-paristech.fr

from different distances and viewpoints in order to characterize their 3D shape. In the following, we will focus on the 3D keypoint detection and the 3D descriptor extraction.

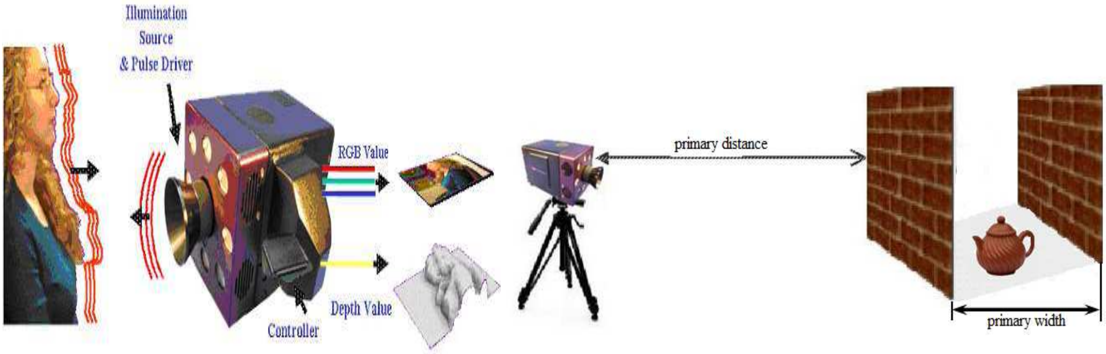

Figure 1. The time-of-flight (TOF) camera used produces depth video with the following principle: measure of time delay between infrared pulse emission and the reception of its reflection.

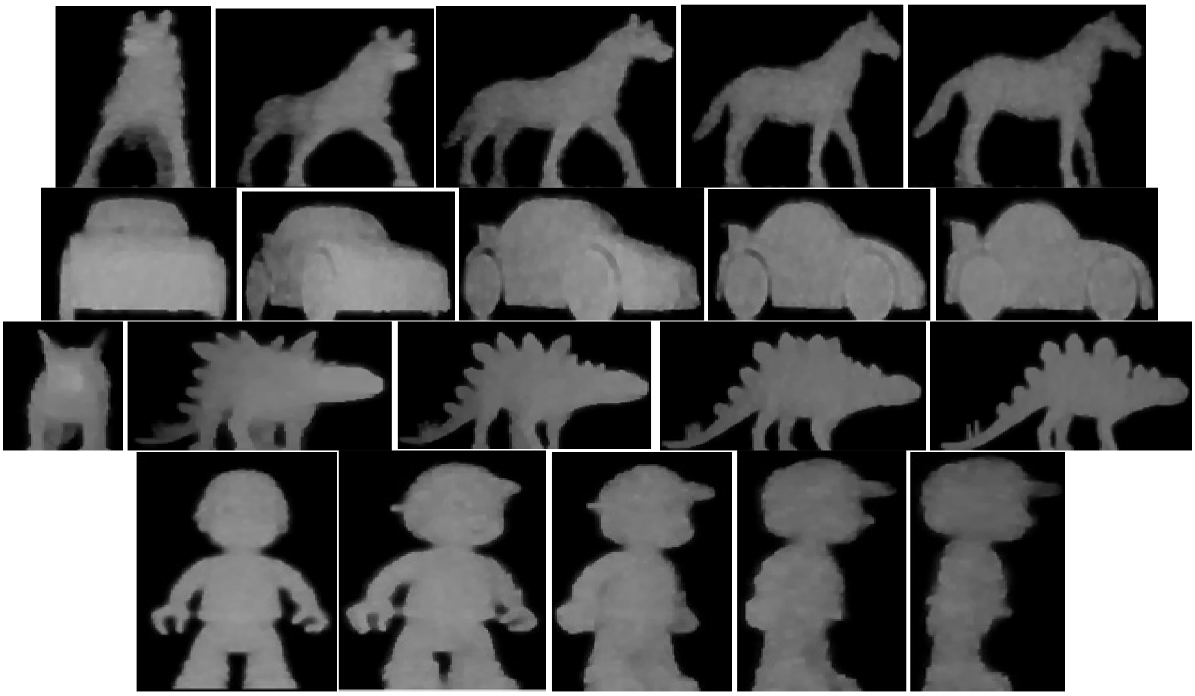

Figure 2. The range images of the four objects of our dataset.

2. METHODOLOGY

2.1 General scheme of the method

Pre-processing

Our methodology is the following: objects are placed on a turning tray which is pivoted to 5 positions (0°, 25°, 50°, 75° and 100°) in front of the camera during each record. Then, we repeat this for 3different distances of the camera to objects (at 50cm, 80cm and 110cm). The 3D camera produces 30 depth images/s, and the typical total recording time is 3.3 seconds, therefore, the total number of frames is fixed to 100 for each record.

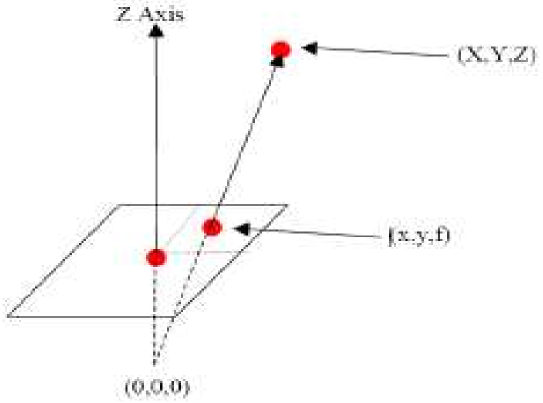

Figure 3: Relation between true 3D point (X,Y,Z), and the (x,y) position of depth pixel in the focal plane.

The camera's output is a depth grayscale image that we convert into a cloud of 3D points. The computation of the 3D points from the depth image is straightforward, as illustrated in Figure 3: for each pixel at line i and column j, we first compute (using the actual pixel width 0.0112 mm on the sensor), its (x,y) position in the focal plane; from this and the focal length f, we can deduce the normalized directing vector $\bar{V} = \frac{(x,y,f)}{\sqrt{x^2 + y^2 + f^2}}$; the true 3D point (X,Y,Z) is then obtained as $(X, Y, Z) = D.\bar{V}$ where the distance D depends directly on the grayscale value g of the depth image by $\bar{D} = P_d + P_w.\frac{255-g}{255}$; where $P_d \; and \; P_w$ are respectively the primary distance (i.e. minimal range distance) and the primary width (i.e. difference between maximal and minimal range distance), which can both be tuned manually on the camera.

One of the problems with TOF depth cameras is the rather high level of noise of the output data. The absolute precision of each depth pixel is ~1cm only (and quite dependant on the reflectance of the object material), and there can be an offset of absolute distance as high as 6cm, according to our tests. We partly overcome the noise problem by averaging the 100 frames recorded during 3s into one depth image that we crop after to remove outliers in boundaries.

Method

Once the 3D points cloud have been generated from the mean frame, we select salient 3D points before computing a 3D descriptor of regions around those detected points. Our method is based on differential geometry to describe the shape of objects. Particularly, we consider surface normals and curvatures which measure how the surface bends in different directions at one point.

In precice, a uniform n x n lattice (grid) is used to sample the 2D cropped depth map (where n = 10). Then, each cell is subdivided into r x r sub-regions (r = 3). Using points belonging to this window, we fit a quadratic surfaces to the r x r patchs, of the form f(x,y) = ax^2 + by^2 + cxy + dx + ey + f, and estimate the parameters of the quadratic surface with the least square method. That allows us to compute differential geometry and extract the surface normal, Gaussian curvature and principal curvatures at each patch. Using a factor quality based on curvatures, we select feature points with the largest shape variation. Then, the shape index values are cumulated and the histogram of angles between the normal of reference feature point and that of its neighbouring regions is computed. Hence, our descriptor is the combination of the two histograms forming a 17+34=51 dimensional vector. To evaluate the proposed detector and descriptor, comparison with Chen's detector and descriptor in term of stability and invariance to the 5 viewpoints and the 3 scales of the same object has been done.

2.2 Keypoint detectors

The first detector is based on a keypoint quality measure introduced by Mian et al. [11]. After the sampling step (100 cells), we associate at each cell k a quality measure Q$_k$ is given by:

$$Q_k = \frac{1000}{l^2} \sum|K| + \max(100K) + |\min\,(100K)| + \max(10\,{k_p}^1) + |\min\,(10{k_p}^2)|\,;\; K = {k_p}^1{k_p}^2 \quad (1)$$

where k^1_p and k^2_p are maximum and minimum principal curvatures, respectively. Summation, maximum and minimum values are calculated over the r x r sub-regions. Absolute values are taken so that positive and negative curvatures do not cancel each other; positive and negative values of curvatures are equally descriptive. Keypoints are ranked according to this measure and a threshold is chosen to select the best ones.

The second detector is the one proposed in [10], and uses the shape index (I_p) for feature point extraction. It is a quantitative measure of the surface shape at a point p, and defined by Eq. (2)

$$I_p = \frac{1}{2} - \frac{1}{\pi}\, arctg\, \frac{k_p^1+k_p^2}{k_p^1-k_p^2} \qquad (2)$$

With this definition, all shapes are mapped into the interval [0, 1] [12]. Larger shape index values represent convex surfaces and smaller shape index values represent concave surfaces.

The central point is marked as a feature point if its shape index I_p satisfies Eq. (3) within an r x r window

I_p = max of shape indexes and $I_p >= (1+\boldsymbol{\alpha}) * \mu$;

or I_p = min of shape indexes and $I_p <= (1- \boldsymbol{\beta}) * \mu$;

where μ is the mean of shape index over the l*l values and $0 <= \boldsymbol{\alpha}, \boldsymbol{\beta} <= 1$ (3)

In Eq. (3) $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ parameters control the selection of feature points.

2.3 Keypoint descriptors

Around each selected keypoint P, a local patch R is constituted of the r x r neighbours. For every point R_i belonging to R, we compute its shape index value and the angle θ between the surface normals at the feature point P and N_i. Then, we form, first, a 2D histogram by accumulating points in particular bins along the two axes based on Eq. (4) which relates the shape index value and the angle to the 2D histogram bin (h_x, v_y). One axis of this histogram is the shape index which is in the range [0, 1]; the other is the cosine of the angle (cos θ) between the surface normal vectors at P and one of its neighbours in R. It is equal to the dot product of the two vectors and it is in the range [-1, 1]. In Eq.(4), (h_x, v_y) are the indexes along the horizontal and vertical axes respectively and (b_x, b_y) are the number of bins along the horizontal and vertical axes, respectively.

$$h_x = S_i \times b_x \; ; \; v_y = \frac{(\cos\theta + 1)\times b_y}{2} \qquad (4)$$

This 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.

The second proposed descriptor concatenates the histogram of the cosines of angles between normals and the histogram of the shape index into 1D vector. The unique axis of the histogram is composed of $b_x + b_y$ bins. (cf. Figure 4).

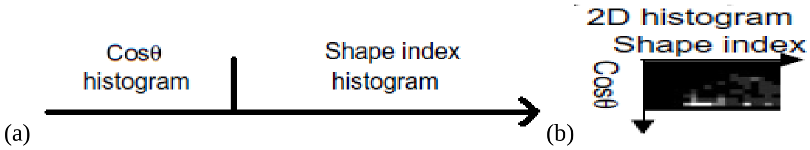

Figure 4. (a) Concatenate descriptor, (b) Combined descriptor

3. EXPERIMENTAL RESULTS

We performed our experiments on our own database constituted of 4 objects (horse, car, dinosaur, man doll) captured with the TOF camera which has a spatial resolution of 320 x 240. The computation time of keypoint detection and description phases is about 2.5s for 50 keypoints. In the following, we propose to evaluate our detector and descriptor in terms of stability and descriptiveness.

3.1 Keypoint stability

In order to measure the repeatability of detected keypoints between different views/scales, we compute the distance of every keypoint in the rotated/scaled point cloud of view 1 to the nearest neighbor keypoint detected in view 2. Figure 5 illustrates the two plots of keypoint repeatability between the four initial views of the four objects and their respective scaled and rotated views with the quality factor based detector (FQD) and the shape index based detector (SID). The y-axis shows the percentage keypoints of the transformed views which could find a corresponding keypoint in the initial view within the distance shown on the x-axis. Results show that the percentage of keypoints repeatability is more important for the FQD than the SID and repeatability reaches 100% at a nearest neighbor error of ~7.5 mm for FQD and at ~10mm for SID. This result suggests that quality factor detector has slightly higher repeatability than the shape index detector.

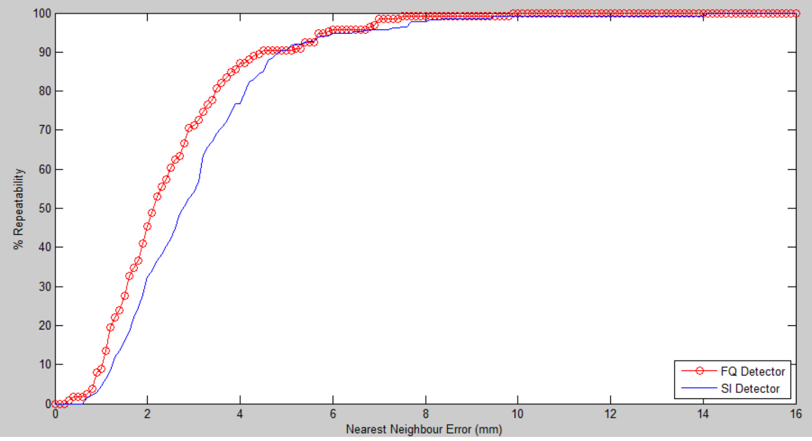

Figure 5. Keypoint identification repeatability between different scales and views for the two detectors: quality factor based detector (FQD) and shape index based detector (SID).

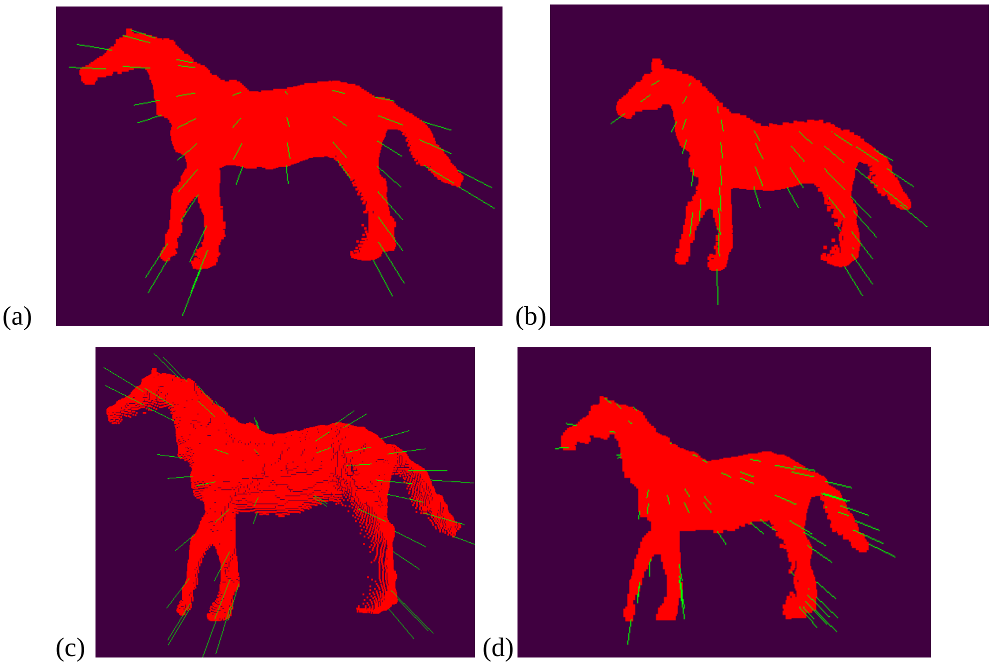

Figure 6. Illustration of the greater stability of Factor Quality Detector (FQD) vs Shape Index Detector (SID , by positions of detected keypoints (shown with green arrows) for two different scales in horse 3D points cloud: top line, detection with FQD, at respective scales 50 (a) and 80 (b); bottom line, detection with SID, at respective scales 50 (c) and 80 (d).

Figure 6 illustrates the relative stability of keypoint's positions when varying scale for the same object.

In addition to the keypoint detection role, the quality measure provides a means of selecting the best required keypoints. A threshold is used to keep the keypoints with Q_k greater than the threshold. Figure 7 shows keypoints detected on one view of horse at different cutoff thresholds of the quality Q_k. Notice that as the threshold is decreased, more and more keypoints appear at less curved parts of the model.

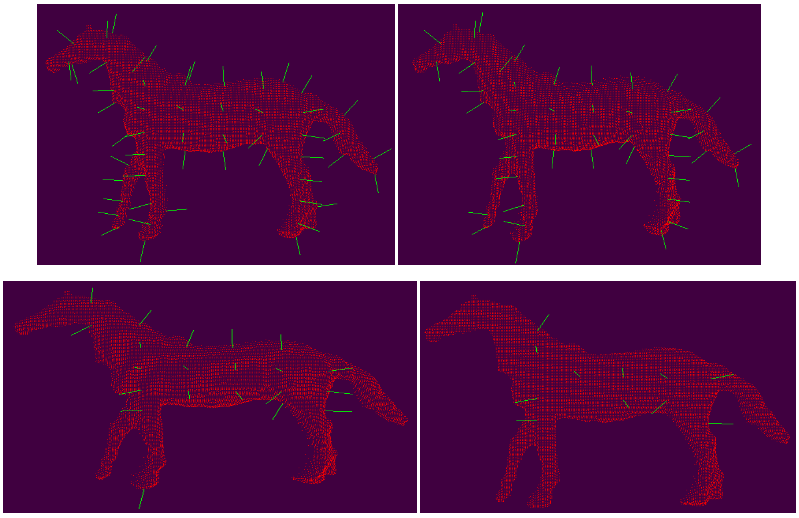

Figure 7. (a)Pointcloud of the horse, $Q_k > 0$, KpNb (the number of keypoints) = 53, (b) For $Q_k > 300$, KpNb = 44
(c) For $Q_k > 7000$, KpNb = 21 (d) $Q_k > 17000$, KpNb = 11

3.2 Descriptor stability

We employ a rank distribution metric for evaluating the stability of a keypoint descriptor model [13]. This measure gauges the probability of finding a matching descriptor in the set of k-nearest neighbors as a function of k. To compute the rank distribution, we fix our match set M of descriptors between descriptors in view 1 and those of view 2. Then, we consider every pair of descriptors (i, j) in the match set M, and count the number of descriptors k in descriptor set of view 1 such that $\|\; i - k \;\|_{L2} < \|\, i - j \,\|_{L2}$.

Results are performed on the point cloud of the four objects at view angles 0°, 25°, 50° and 100° and at scales 50, 80 and 110.

After selecting n points and computing their descriptor in view 1, rotation of the n points is applied in order to recover their position in view 2 and compute their descriptor. Thus, the descriptor stability could be evaluated by counting the number of points in view 2 which the closest descriptor distances corresponds to the initial point in view 1.The same process is done for scale. We compute the positions of keypoints in scale 2 corresponding to the detected keypoint in scale 1 by applying a homothetic transformation. We plot the mean of curves representing the evolution of the number of correct matching keypoint descriptor as a function of the number of K nearest descriptors.

As illustrated in figure 8, according to this metric, the combined model slightly outperforms the concatenate one for most values of K, including small ones which are the most relevant.

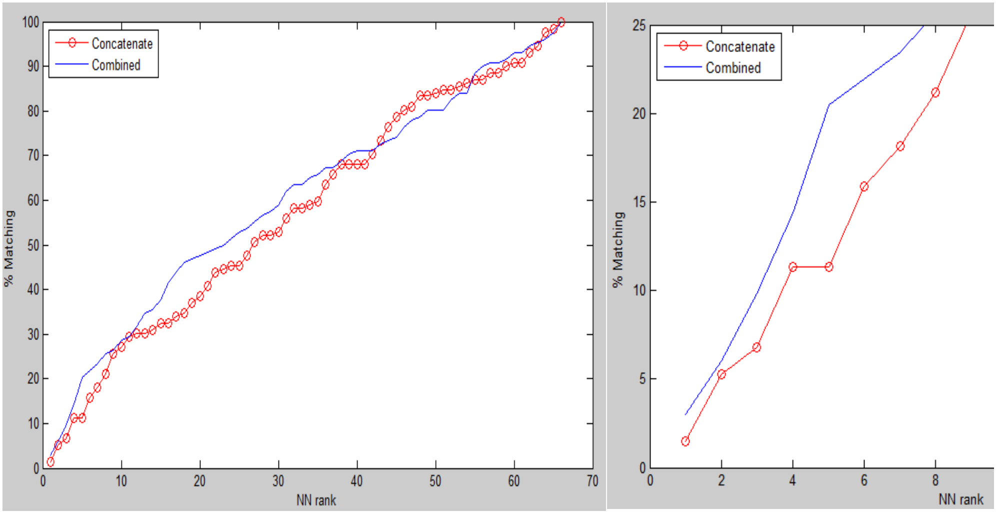

Figure 8. Rank distribution for each of the two descriptor models (concatenate, combined). The plot shows the probability of finding the correct match for a descriptor within the group of the k nearest neighbors in the initial keypoint set. Right plot is a zoom on region for small K values.

3.3 Descriptiveness of the descriptors

In order to recognize one object appearing at different orientations or scales, the used descriptors must keep quite similar values whatever the considered view. Let E_j be the vector formed with values $\sigma_{1,j}$, $\sigma_{2,j}$, $\sigma_{3,j}$, $\sigma_{4,j}$, where $\sigma_{i,j}$ is the standard deviation of the jth descriptor values obtained for the ith object. To obtain close values for the same object, the mean value $,\overline{E_j}$ of E_j , must be minimized. In order to discriminate objects, the used descriptors must present quite different values to describe separate objects. Let M_j be the vector formed with values $m_{1,j}$, $m_{2,j}$, $m_{3,j}$, $m_{4,j}$, where $m_{i,j}$ the mean value of the jth descriptor values obtained for the ith object. To discriminate objects, the standard deviation of M_j , σ_{M_j} , must be maximized. Consequently, the quotient $\sigma_{M_j} / \overline{E_j}$, which have been used in [14], is a criterion which can characterize the descriptor performance.

Table 1 presents the concatenate and the combined descriptor performance regarding to the previously mentioned criterion for four objects of the database. This quotient is slightly more important when using the combined descriptor.

Table 1. Descriptor performance

	Concatenate descriptor	**Combined descriptor**
Quotient ($\sigma_{M_j} / \overline{E_j}$)	5.4148e-005	6.3784e-005

The other criterion proposed to compare the distinctive power of the first representation with the second one is the coefficient of variation which is equal to the standard deviation divided by the mean [15]. The coefficient of variation encodes variability relatively to the mean and is used to compare the relative dispersion in one type of data with the relative dispersion in another type of data.

The diagrams of Figure 9 compare the dispersion of the two tested descriptors. We note that in high values of coefficient of variation the combined descriptor curve is above the concatenate descriptor curve, which supports the conclusion that the combined descriptor is more descriptive than the concatenated one.

One reason to explain this result is the resolution of histogram bins in the combined descriptor which is higher (17*34 = 578) than the concatenate descriptor one (17+34=51).Note that increasing too much this resolution will not necessary improve the results; the models will just fit more noise if each dimension is not supported by a reasonable portion of data points.

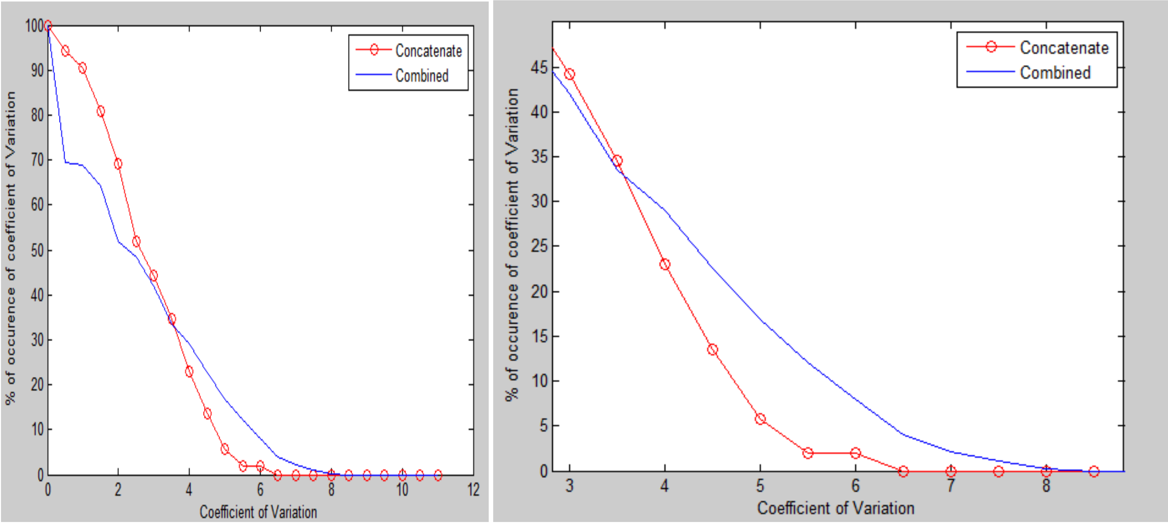

Figure 9. Average of the coefficient of variation for the two representations (concatenate and combined) for the four objects.

4. CONCLUSIONS AND PERSPECTIVES

We originally assumed that computing curvatures on the point would have allowed us to obtain stable 3D keypoints which describes parts of object with noticeable shape variation. Experimentations reported here showed this assumption as relevant. We initially believed that using the shape index would make it possible to encode the convexity and concavity of the surface. We therefore proceeded with our tests to confirm this assumption. It has often been suggested that combining shape index and angles of normals would make it possible to form an invariant descriptor. Trials carried out to test this assumption proved such to be the case.

The experiments presented here indicate higher stability of 3D keypoints selected with a new quality criteria based on curvature under viewpoint variation. Undergoing experiments also indicate higher stability under scale variations. Regarding the descriptor, our analysis suggest that combined ShapeIndex-NormalAngles has more stability and descriptiveness than the concatenate version.

This new proposed 3D keypoint detector and the combined ShapeIndex-NormalAngles 3D descriptor, therefore, has the required properties to allow correct recognition of 3D objects whatever their pose, and distance, thus helping to provide semantic meaning to a complex scene. As we have seen, this result can be explained by the use of differential geometry which permits us to describe the local variation of the surface. It is also likely that if we try combining other 3D descriptors, descriptiveness will be improved. Nevertheless, attention should be paid to the computation time cost for best match searching that could be induced by high dimensionality of descriptor. Forthcoming investigations include tests of our approach on more objects, including public databases, and verification of the performance of these 3D keypoints as a tool for object recognition and categorization.

REFERENCES

[1] Johnson, A.E. and Hebert, M., "Using spin images for efficient object recognition in cluttered 3Dscenes," IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(5), 433-449(2002).

[2] Samir, C. Daoudi, M. and Srivastava, A. "Reconnaissance de Visages 3D Utilisant l'Analyse de Formes des Courbes Faciales," 10èmes Journées CORESA (COmpression et REprésentation des Signaux Audiovisuels), 9-10(2006).

[3] de Diego, I.M. Serrano, A. Conde, C. and Cabello, E. "Face verification with a kernel fusion method,"

Pattern Recognition Letters, 31(9), 837-844 (2010).

[4] Bowyer, K.W. Chang, K. and P. Flynn, "A survey of 3D and multi-modal 3D+ 2D face recognition," Notre Dame Department of Computer Science and Engineering Technical Report, (2004).

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

[6] Li, X. and Guskov, I. "Multi-scale features for approximate alignment of point-based surfaces," in

Proceedings of the third Eurographics symposium on Geometry processing, 217(2005).

[7] Jones, M. and Viola, P. "Face recognition using boosted local features," in Proceedings of international conference on computer vision, (2003).

[8] David G. Lowe, "Object recognition from local scale-invariant features," Proceedings of the International Conference on Computer Vision, 2, 1150-1157(1999).

[9] Lo, T.W.R. and Siebert, J.P. "SIFT keypoint descriptors for range image analysis," Annals of the BMVA, 1-17(2008).

[10] Chen, H. and Bhanu, B. "3D free-form object recognition in range images using local surface patches," Pattern Recognition Letters, 28(10), 1252-1262 (2007).

[11] Mian, A. Bennamoun, M. and Owens, R."On the Repeatability and Quality of Keypoints for Local

Feature-based 3D Object Retrieval from Cluttered Scenes," International Journal of Computer Vision, 89 (2), 348-361(2010).

[12] Dorai, C., Jain, A. "COSMOS—A representation scheme for 3D free-form objects," IEEE Trans. Pattern Anal. Machine Intell. 19 (10), 1115-1130(1997).

[13] Bosse, M. and Zlot, R. "Keypoint design and evaluation for place recognition in 2D lidar maps," Robotics and Autonomous Systems archive, 57(12), 1211-1224 (2009).

[14] Choksuriwong, A. Laurent, H. and Emile, B. "A Comparative Study of Objects Invariant Descriptor," ORASIS, (2005).