EurographicsWorkshop on 3D Object Retrieval (2010)
I. Pratikakis, M. Spagnuolo, T. Theoharis, and R. Veltkamp (Editors)

3D Object Recognition and Facial Identification Using Time-averaged Single-views from Time-of-flight 3D Depth-Camera

Hui Ding, Fabien Moutarde, and Ayet Shaiek

Robotics Lab (CAOR), Mines ParisTech, Paris, France
Fabien.Moutarde@mines-paristech.fr

Abstract

We report here on feasibility evaluation experiments for 3D object recognition and person facial identification from single-view on real depth images acquired with an "off-the-shelf" 3D time-of-flight depth camera. Our methodology is the following: for each person or object, we perform 2 independent recordings, one used for learning and the other one for test purposes. For each recorded frame, a 3D-mesh is computed by simple triangulation from the filtered depth image. The feature we use for recognition is the normalized histogram of directions of normal vectors to the 3D-mesh facets. We consider each training frame as a separate example, and the training is done with a multilayer perceptron with 1 hidden layer. For our 3D person facial identification experiments, 3 different persons were used, and we obtain a global correct rank-1 recognition rate of up to 80%, measured on test frames from an independent 3D video. For our 3D object recognition experiment, we have considered 3 different objects, and obtain a correct single-frame recognition rate of 95%, and checked that the method is quite robust to variation of distance from depth camera to object. These first experiments show that 3D object recognition or 3D face identification, with a time-of-flight 3D camera, seems feasible, despite the high level of noise in the obtained real depth images.

Keywords: 3D face identification, 3D object recognition, time-of-flight camera, depth image, range data

1. Introduction

Person identification by its face, as well as object recognition, can work very well for a particular viewpoint (see for instance [ZCPR03] and [PM*00] for a survey of face recognition methods and performances, and [MG06] for a more general overview of visual object recognition techniques (focusing on the difficult case of pedestrian recognition). But achieving robustness to pose or viewpoint variations for these applications is still a quite challenging problem (see e.g. [ZLWW07][TFL*06][LHB04]). Meanwhile, new devices are appearing, such as time-of-flight (TOF) 3D cameras, which can make it possible to use 3D data rather than classical 2D images, for face identification and object recognition. And more and more work are published these last years on using depth images for 3D object recognition. In [HLLS01], Hetzel et al. compare the interest for 3D object recognition from range images of 3 different features: depth histograms, normal histograms and curvature histograms. In [TM07], Tsalakanidou and Malassiotis are directly using the depth map as an image, from which DCT coefficient are computed and used as observation for an Embedded Hidden Markov Model (EHMM) for face classification. Mian et al. [MBO07] have also proposed an efficient hybrid approach combining 2D and 3D information. As highlighted in [CF01], the most widely used type of approach is using a global 3D model, and basing the recognition on the comparison of estimated model with reference models, or generate from the 3D model many synthetic 2.5D views, and compare to them the 2.5D scans to be identified (see eg the work of Lu et al. in [LJC06]). Some teams also explore more local methods, as Savarese and Fei-Fei in [SF07], where they propose a "3D-part" based model for 3D object matching. Other teams use specific global 3D appearance features, such as shapes of level curves of the depth image for faces in [SSD06].

In this paper, we report on investigations conducted in our lab on the feasibility of doing real-time 3D face identification and 3D object recognition using the depth video of a time-of-flight 3D camera.

2. Experimental set-up

2.1. 3D camera

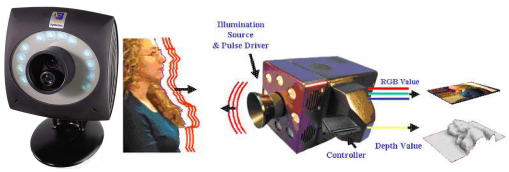


Figure 1: *The time-of-flight (TOF) camera used produces depth video with following principle: measure of time delay from infrared pulse emission to the reception of its reflection*

The device we use is the z-cam by 3DVsystems, which produces 320 × 240 depth images at 30 frames per seconds. The range-sensing technology of this camera uses an illumination source that sends out pulsed infrared signals, and a fast gating & timing unit. The pulses are reflected by the objects in the scene, and the depth sensor then measures the accumulation of photons to determine the exact distance

of each pixel in the scene. The range data is coded as a grayscale video, with grey level proportional to distance.

2.2. Methodology and pre-processing

Our methodology is the following: for each person or object, we perform 2 *independent* recordings, one for learning and the other for test; in the case of objects, they are placed on a turning tray which is pivoted 360 degrees in front of the camera during each record; for persons, only the face is in the camera field, and the person moves head slowly 90 degrees to the left, then back to center, and finally 90 degrees to the right. The 3D camera produces 30 depth images/s, and the typical total recording time is 15-20 seconds, therefore, the total number of frames is ~ 500 for each record.

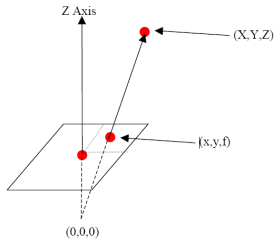


Figure 2: *Relation between true 3D point (X,Y,Z), and the (x,y) position of depth pixel in the focal plane.*

Since we want to use features derived from the 3D surface of the object, we need to convert the depth grayscale image into a cloud of 3D points, and then compute a 3D-mesh for these points. The computation of the 3D points from the depth image is straightforward, as illustrated in figure 2: 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 $\vec{V} = \frac{(x, y, f)}{\sqrt{x^2 + y^2 + f^2}}$; the true 3D point (X,Y,Z) is then obtained as $(X, Y, Z) = D \cdot \vec{V}$ where the distance D depends directly on the grayscale value g of the depth image by $D = P_d + P_w \cdot \frac{255 - g}{255}$ 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 (see 3rd image of figure 3). 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 applying, for each record frame, once the 3D points are computed, a median filtering for reducing noise. A 3D-mesh is then built by a simple triangulation linking only adjacent pixels, as shown on left of figure 4. These three successive processing steps are illustrated on figure 3, on which one can see the grayscale depth image, the corresponding cloud of 3D points, and finally the 3D-mesh without or with preliminary median filtering of the 3D points.

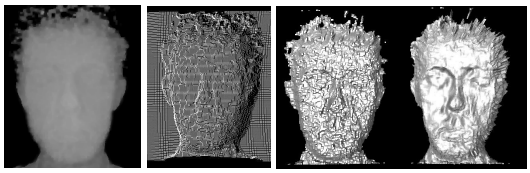


Figure 3: *Typical depth image (left), cloud of 3D points (2nd left), and 3D-mesh before and after filtering (right).*

2.3. Feature extraction: histogram of normal vectors

As mentioned in [HLLS01], one could think of using directly histogram of pixels depth. These are invariant under translations and image plane rotations, and also to scale if distance is normalized. However, normalized distances are very sensitive to the perceived depth range. Another problem with distance histograms is that they may be influenced by neighboring objects or background clutter.

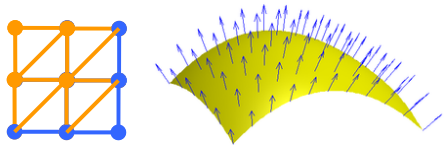
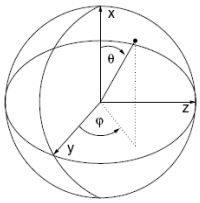


Figure 4: *On the left, the simple triangulation used for building the 3D-mesh: adjacent pixels are linked; on the right, illustration of regularly spaced surface normals, from which we compute a normalized normals orientation histogram.*

Therefore, considering also the results presented in [HLLS01], and computation time considerations, we chose as our feature for 3D object recognition the normalized histogram of directions of normal vectors to the 3D-mesh. This feature may be thought as weak and not enough discriminative, but it has the great advantage of being theoretically intrinsically invariant under small (relative to distance) translations of the objet in the view-field, and under moderate relative variation of distance to the object.



$$\phi = \arctan\left(\frac{n_z}{n_y}\right), \qquad \theta = \arctan\frac{\sqrt{n_y^2 + n_z^2}}{n_x}$$

Figure 5: *Illustration of angles used to measure the orientation of normals, and the formulas to compute them from the normal vector.*

The direction is characterized by inclination angle θ and azimuth angle ϕ, both in [0; π[, and easily computed from the normal vector, as illustrated on figure 5. For each angle, the [0; π[interval is subdivided in 8 bins, so that the histogram has a total of 8x8=64 bins, and the histogram is normalized (i.e. for each of the 64 bins, we compute the *proportion* of normals falling into it), so it is a vector of

[0;1[^64. We consider each frame as a separate example, and each one is thus represented by a 64-dim normalized vector.

3. Recognition results

3.1. 3D facial identification

For our 3D person identification experiments, 3 different persons were used for our first preliminary experiment. We first tried for the recognition a simple best-match approach using the minimal $\chi 2$ divergence (as proposed by [HLLS01]) where the latter is given, for 2 histograms (q_i) and (v_i), by:

$$\chi^2(Q,V) = \sum_i \frac{(q_i - v_i)^2}{q_i + v_i}$$

The correct identification rate (i.e. rank-1 recognition rate) in our experiment is 76%, as shown in table I.

TABLE I. *FACE CORRECT IDENTIFICATION RATES (RANK-1 RECOGNITION) WITH HISTOGRAM MATCHING USING $\chi 2$ DIVERGENCE*

Base / Test	Chris (%)	Hui (%)	Raoul (%)
Chris	67	3	30
Hui	7	70	23
Raoul	2	8	90
Mean	76 %		

However, this first method has an important drawback: the computation time is proportional to the number of classes, and worse, to the number of reference histograms for each class. Another problem is that is does not allow easy implementation of a "reject" policy for not classifying ambiguous cases. And finally, it requires to store in memory a large number of reference histograms. We therefore propose classification method based on the same normalized histogram of normal orientations: training a multi-layer neural network, with as input the 64-dim normalized histogram, and as many output as the number of classes (1 versus all classification encoding). The ~500 histograms for each class computed from the frames of the training records are used as training set, and the correct recognition results are evaluated on the normalized histograms computed from the frames of the test records. The hidden layer size was chosen as 7, as the smallest value giving the best result among 4 sizes tested (3, 7, 10, 15). The global correct recognition rate is similar, although a bit lower: 72%, as can be seen on table II.

TABLE II. *FACE CORRECT IDENTIFICATION RATE RATES (RANK-1 RECOGNITION) USING A MULTI-LAYER NEURAL NETWORK CLASSIFIER APPLIED TO NORMALIZED HISTOGRAM OF ONE SINGLE DEPTH FRAME*

Base / Test	Hui (%)	Raoul (%)	Chris (%)	Not classified (%)
Hui	56	21	5	18
Raoul	4	89	1	6
Chris	2	21	71	9
Mean	72 %			

In order to further improve our recognition rates, we tried to apply the same method on *histograms averaged over several successive depth frames*. As can be seen on figure 6, the global correct identification rate significantly increases with increasing number of successive frames used for averaging, and can reach up to 80%. This is an illustration of the importance of filtering data from TOF camera, as averaging over several frames is equivalent to some temporal filtering. However if we also consider separate recognition rate for each person, averaging on only 3 frames seems optimal. The corresponding confusion matrix is shown in table III. Comparison with table II shows considerable improvement, and this last result is also globally better than that of table I.

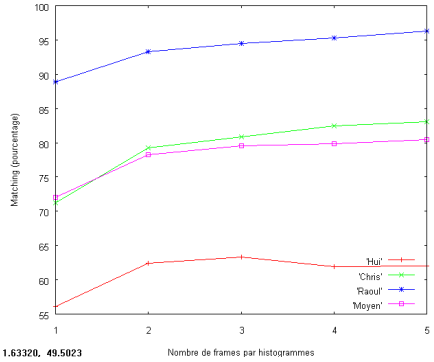


Figure 6: *Influence on correct recognition rates of the number of successive depth frames used for averaging histograms.*

TABLE III. *FACE CORRECT IDENTIFICATION RATE USING A MULTI-LAYER NEURAL NET CLASSIFIER APPLIED TO NORMALIZED HISTOGRAM AVERAGED OVER 3 SUCCESSIVE DEPTH FRAMES*

Base / Test	Hui (%)	Raoul (%)	Chris (%)	Not classified (%)
Hui	64	16	4	16
Raoul	2	94	1	3
Chris	1	12	80	7
Mean	79 %			

3.2. 3D object recognition

We have conducted similar preliminary experiments for 3D object recognition. We have used three quite dissimilar objects: a soft rubber toy giraffe, a hard plastic robot dog, and a clay tea-pot. Figure 7 shows examples of depth images and 3D-meshes for these 3 objects. It can be noticed that 3D-mesh seems more accurate for the tea-pot. We applied exactly the same method as for the face identification, but only to histograms computed on single depth frames. The global correct rank-1 recognition rate is a very high 95%, which is rather natural given the high dissimilarity of the three objects. For comparison, the recognition rate reported in [HLLS01] is 80% using $\chi 2$ best-match of normals histogram for a larger set of 30 objects, but with artificial perfect depth images, while we use real depth images, which are rather noisy.

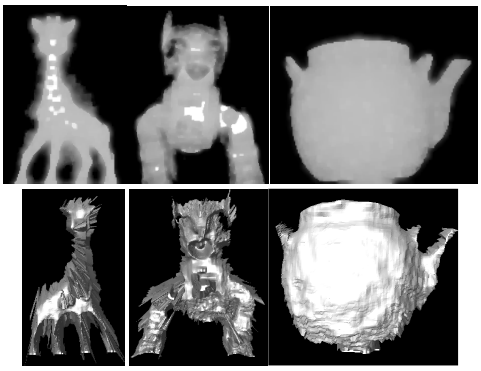


Figure 7: *Examples of depth images (top line) and computed 3D-mesh (bottom line) of objects used for 3D object recognition.*

TABLE IV. *OBJECT CORRECT RANK-1 RECOGNITION RATE USING A MLP APPLIED TO NORMALIZED HISTOGRAM OF SINGLE DEPTH FRAME*

Base / Test	Giraffe (%)	Robot dog (%)	Tea pot (%)	Not classified (%)
Giraffe	**92**	**3**	**0**	5
Robot dog	**3**	**93**	**0**	4
Tea pot	**0**	**0**	**99**	1
Mean	**95 %**			

3.3. Robustness to distance variations

In order to estimate the robustness of the 3D object recognition to variations of distance from the depth camera to the object, we have made 2 more test recordings of the Giraffe object, one at longer distance (~65 cm) and one at shorter distance (~25 cm) than the initial distance (~45 cm) used for the training and first testing records of the same object. As reported on table V, the Giraffe recognition rate is only very slightly degraded when distance is significantly different from the distance used in the training set: the correct recognition rate remains above 89% at shorter and longer distances.

TABLE V. *INFLUENCE OF DISTANCE VARIATION ON OBJECT RECOGNITION*

Base / Test	Giraffe (%)	Robot dog (%)	Tea pot (%)	Not classified (%)
Giraffe 25cm	**89**	**3**	**0**	8
Giraffe 45cm	**92**	**3**	**0**	5
Giraffe 65cm	**90**	**5**	**0**	5

4. Conclusions and perspectives

We have presented our first experiments for 3D object recognition and person facial identification with a time-of-flight 3D camera. Using normalized histogram of directions of normals to the 3D-mesh as feature, and a simple multilayer neural network as a classifier, we obtained a global correct rank-1 recognition rate among faces of 3 different persons of up to 80%, when averaging histograms over several frames. Using the same approach on more dissimilar objects, we get a single-frame rank-1 recognition rate of 95% on single-frame histograms. Even though these results are for very small number of person/objects, and further experiments (currently in progress) on larger sets, and with more dissimilar objects, are obviously necessary to draw stronger conclusions, we can already conclude from these first experiments that, despite the high level of noise in the obtained real depth images, 3D object recognition with our current approach seems feasible for objects with sufficiently different 3D shapes, and 3D person recognition probably requires to be done on sequences larger than 1 frame to attain sufficient recognition rates. Other perspectives include optimization of the computation of normals from the 3D points, or investigation of other features, in order to approach real-time recognition.

References

[CF01] CAMPBELL R. J., FLYNN P. J., "A survey of free-form object representation and recognition techniques", Computer Vision and Image Understanding, Volume 81, Issue 2, pp. 166 – 210, 2001.

[HLLS01] HETZEL G., LEIBE B., LEVI P., SCHIELE B.: "3D Object Recognition from Range Images using Local Feature Histograms", proc IEEE *conference on Computer Vision and Pattern Recognition (CVPR'01), 2001.*

[LHB04] LECUN L Y., HUANG F.J., BOTTOU L.: "Learning methods for generic object recognition with invariance to pose and lighting", *Proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR'04)*, Washington D.C., USA, 2004.

[LJC06] LU X., JAIN A.K., COLBRY D.: "Matching 2.5D face scans to 3D models", IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI), 28(1), pp. 31-43, Jan. 2006.

[MBO07] MIAN A. S., BENNAMOUN M., OWENS R.: "An efficient Multimodal 2D-3D hybrid approach to automatic face recognition", IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI), 29(11), pp. 1927-1943, November 2007.

[MG06] MUNDER S., GAVRILA D. M.: "An Experimental Study on Pedestrian Classification". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.28, nr 11, pp. 1863-1868, 2006.

[PM*00] PHILIPS P.J., MOON H., et al.: "The FERET Evaluation Methodology for Face-Recognition Algorithms", *IEEE Transactions on PAMI*, Vol.22, No.10, pp1090-1104, 2000.

[SF07] SAVARESE S., FEI-FEI L.: "3D generic object categorization, localization and pose estimation", IEEE Intern. Conf. in Computer Vision (ICCV), 2007.

[SSD06] SAMIR C., SRIVASTAVA S., DAOUDI M.: "Three-dimensional face recognition using shapes of facial curves", IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI), 28(11), pp. 1858-1863, 2006.

[TFL*06] THOMAS A., FERRARI V., LEIBE B., TUYTELAARS T. , SCHIELE B., VAN GOOL L.: "Towards Multi-View Object Class Detection", *Proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, USA, 2006.

[TM07] TSALAKANIDOU F., MALASSIOTIS S.: "Application and evaluation of a 2D+3D face authentication system", 3DTV Conference, 2007.

[ZCPR03] ZHAO W. , CHELLAPA R., PHILIPS P.J., ROSENFELD A.: "Face Recognition: A Literature Survey" , ACM Computing Surveys (CSUR), Volume 35, Issue 4, pages: 399 – 458, December 2003.

[ZLWW07] ZHANG H., LI Y., WANG L., WANG C.: "Pose insensitive Face Recognition Using Feature Transformation", IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.2, February 2007.