21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Real-Time 3D Face Identification from a Depth Camera

Rui Min
EURECOM-France
min@eurecom.fr

Jongmoo Choi
University of Southern California-USA
jongmoo@usc.edu

Gérard Medioni
University of Southern California-USA
medioni@usc.edu

Jean-Luc Dugelay
EURECOM-France
jld@eurecom.fr

Abstract

We present a real-time 3D face identification system using a consumer level depth camera (PrimeSensor). Our system takes a noisy sequence as input and produces reliable identification. Instead of registering a probe to all instances in the database, we propose to only register it with several intermediate references, which considerably reduces processing, while preserving the recognition rate. The presented system routinely achieves 100% identification rate when matching a (0.5-4 seconds) video sequence, and 97.9% for single frame recognition. These numbers refer to a real-world dataset of 20 people. The methodology extends directly to very large datasets. The process runs at 20fps on an off the shelf laptop.

1. Introduction

Most existing 3D face recognition systems use laser scanning [1], stereo vision [2] or structure from motion [3] to obtain the 3D face models. Unfortunately, a laser scanner is slow and expensive, and multi-images approaches suffer from costly processing. Thanks to the recent success of low-cost RGB-D cameras, such as PrimeSensor/Kinect, depth information is directly provided by the sensor. These sensors have recently received attentions from many researchers, especially in the fields of Human-Computer Interaction [4] and 3D object modeling [5].

We want to develop a fast and accurate online 3D face identification system based on the acquisition of the 3D face data from a low-cost depth camera (the PrimeSensor), since it is the interest of many entertainment and security applications. However, in comparison to the 3D faces in many standard dataset (e.g. the FRGC database [1]), a 3D face obtained from the PrimeSensor is rather noisy and incomplete, at a relatively low resolution (640×480). The proposed system overcomes these limitations via a number of processing steps and a robust face registration architecture.

One of key issues in 3D face recognition is to align 2 face shapes in a way that they can be better compared, namely face registration. Iterative Closest Point (ICP) methods [6][7] are the dominating techniques for 3D face registration, since the first work presented by Medioni and Waupotitsch [2]. According to a recent review by Spreeuwers [8], face registration in the literature can be categorized into 3 classes: (1) One-to-all registration (e.g. [9]), where a probe face is registered to all gallery faces, (2) Registration to a face model or atlas, typically to an Average Face Model (AFM) [10], (3) Registration to an Intrinsic Coordinate System (ICS) [8]. The one-to-all approach is known to be accurate but slow in the identification mode since ICP is a time-consuming process. Registration to AFM or ICS has an indirect registration architecture, which requires only 1 time ICP process during an online query. However, the generation of AFM and ICS relies on the good land-marking of gallery faces, which is difficult for the noisy, incomplete and low-resolution faces as in our case.

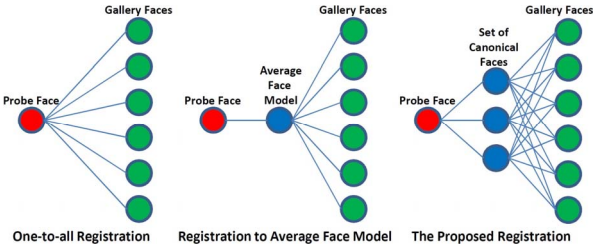

Fig. 1: Architecture of one-to-all registration (left), architecture of registration to a AFM or ICS (middle), architecture of the proposed registration (right).

Here, we propose to apply indirect face registration via multiple references (a small number of canonical faces), and we register both the probe face and gallery faces to this set of canonical faces (see Fig.1), for the following reason: during an online query, the registration needs only a few ICP processes (e.g. 3-5); there is no effort for land-marking; we demonstrate via experiments that increasing the number of canonical faces can significantly improve the identification results. For the fast computation of ICP, we adopted the EM-ICP [11] algorithm based on a GPU implementation [12]. Our system can then work in real-time, with average speed ranging from 0.04s to 0.38s (depending on the number of canonical faces used).

* This research is funded in part by HP Labs Innovation Research Program (CW 218094). EURECOMs research is partially supported by its industrial members: ORANGE, BMW Group, Swisscom, Cisco, SFR, ST Ericsson, Symantec, SAP, and Monaco Telecom.

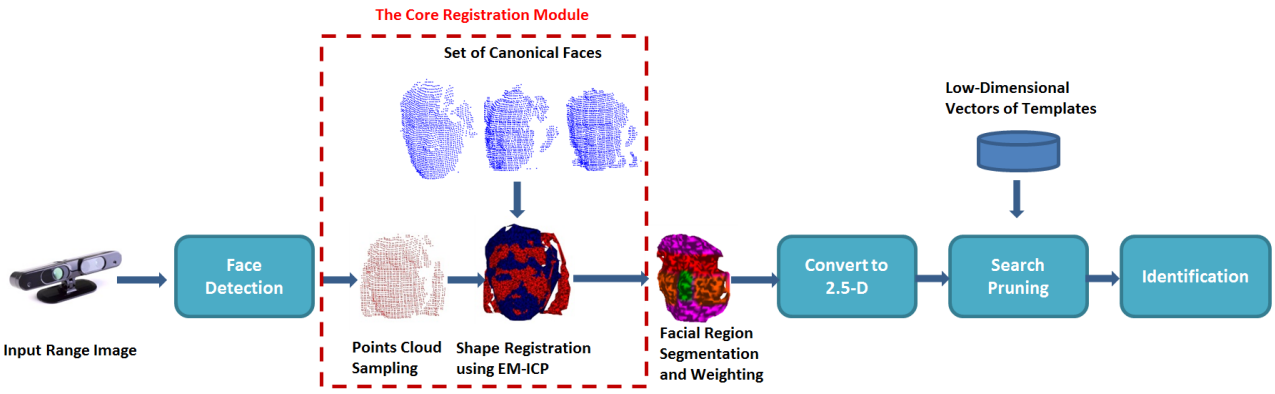

Fig. 2: System overview.

Contributions:

(1) We present a complete real-time 3D face identification system using a depth camera.

(2) A face registration architecture using multiple intermediate references.

(3) Validation on a real-world 20 people dataset.

The rest of this paper is structured as follows. The details of the proposed method are described in section 2. In section 3, extensive experiments are shown to justify our approach. Finally we draw the conclusion in section 4.

2. Method

The proposed system implements a sequence of processing steps including: face detection and segmentation, the core registration module, facial region segmentation and weighting, 2.5D conversion, search space pruning and finally the one-to-many matching. A comprehensive overview is given in Fig. 2.

2.1. Face detection and segmentation

The output from the PrimeSensor includes a RGB image and a depth map at 640×480 resolution. Although the face detection could be achieved by the popular Viola-Jones' method [13] using RGB images, it cannot segment the face/head region exactly from the background/body part. In addition, RGB images are sensitive to illumination variations. Therefore, we focus on the depth information from the range camera (which is illumination-invariant). Because pixels on the head surface have close depth values, given a pre-defined threshold, it is easy to segment the head region according to the depth discontinuity.

Given the segmented visible surface of a head, we first subsample points at a fixed resolution (60×60 in our system) and then compute corresponding real-world 3D coordinates. Any face with a lower resolution is automatically rejected as an invalid face candidate.

2.2. Core registration module

The one-to-all registration method is too computationally complex for real-time identification. We therefore propose a simple, efficient and robust face registration strategy, which demands only a few ICP processes during an online query and does not require additional efforts of land-marking as for the registrations to AFM or ICS.

First, we randomly select M faces (selected from the gallery) to form a set of canonical faces (instead of random selection, gallery face clustering can also be adopted to select representative reference faces). During offline enrollment, each gallery face g_i ($i \in [1, N]$, where N is the size of gallery) is aligned with the M canonical faces (using ICP), and thus generates a set of aligned gallery faces $\{g'_{i,k}, k \in [1, M]\}$. During an online query, a probe face p is also aligned with the same M canonical faces, and thus generates a set of aligned probes $\{p'_k, k \in [1, M]\}$. An illustration of the alignment of a face to multiple canonical faces is shown in Fig. 3. Later in the matching phase, an aligned probe face is matched with the aligned gallery face which is aligned with the same canonical face. Since the recognition capability of the proposed system relies on the fact that registration of different 3D faces from the same identity yields the same or very similar alignment results, increasing the number of canonical faces for multiple alignments improves our system accuracy.

The major computational burden lies on the registration of all gallery faces (which needs $N \times M$ times alignments), but this is done during offline training. An online query requires only M alignments ($M \ll N$). The proposed registration strategy greatly reduces the computation for online identification.

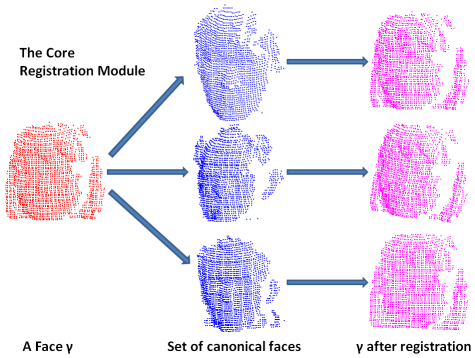

Fig. 3: 3D face registration to a set of canonical faces.

A generic ICP algorithm is time-consuming (it takes several seconds for a typical face data [8]). Here, we adopt an improved version of ICP: the EM-ICP algorithm [11] for face alignment due to its reported accuracy and efficiency. The implementation of the EM-ICP algorithm is provided by Tamaki et al. [12] on CUDA architecture [14] using GPU computing, which is 60 times faster than the OpenMP based implementations on a multi-core CPU.

2.3. Facial region segmentation and weighting

In order to exclude the unstable features (e.g. hair and boundary aliasing), we segment only the facial region. First, the coordinates of the nose-tip is manually annotated for each canonical face. After registration, we suppose the nose-tip of a probe face is aligned with the nose-tip of a canonical face. Then facial region segmentation is done by masking the registered probe face based on a Euclidian distance from the nose-tip.

We further divide the facial region into different facial areas (nose, eyes region, cheeks region and the rest part) [10]. Each area is associated with a weight ({3, 2, 1, 0.5}) to indicate its importance in face identification. The choice of these weights is based on experiments on a validation set.

2.4. Convert from 3D to 2.5D representation

Matching two 3D points clouds demands looking for "hard" correspondence between sets (each point in one set has a unique mapping to a point in the other set). In the case of one-to-all registration, distances are returned directly by the ICP processes. However, when the probe face and all gallery faces are not directly registered, the distance between a probe and a gallery face requires explicit points indexing. One solution is to construct a K-D tree [15] for each points cloud, which is computational expensive in terms of both construction and query.

To reduce the computation of matching two 3D faces, we convert the registered 3D points cloud into a 2.5D representation via orthographic projection. During matching, the pixel-wise comparison of two 2.5D images does not need any explicit indexing and thus much faster than the matching of two 3D points clouds. These 2.5D images (after the proposed face registration) are good enough to establish identity, without further feature extractions on top of them (as shown in [8][10]).

2.5. Vector based search pruning

A brute force comparison (even for the simple 2.5D image matching) of the query and the entire gallery may become extremely computational expensive, especially for large databases. The computational complexity is proportional to the number of entries in the gallery. We propose a two-steps non-linear process to first prune the search space using low-dimensional vectors. One way to compute such a vector for a probe face is to compute the L2 distances between each facial area on the probe face and the corresponding areas on the canonical faces; those computed distances are formed as the vector. The vectors of template faces are computed during the offline training. In this way, we could roughly narrow the search space more than 60% with 100% confidence in our experiments.

2.6. Identification

Given the set of a probe face p registered to M canonical faces $\{p'_k, k \in [1, M]\}$ (please notice that p'_k, $g'_{i,k}$ in section 2.2 are 3D points clouds, whereas here they are range images after the 2.5D conversions described in section 2.4) and the pruned search space $\{g'_{i,k}, i \in [1, N'], k \in [1, M]\}$ (N' is the number of gallery faces in the pruned search space), we find the probe's identity by the following equation:

$$id(p) = \underset{i=1\ldots N'}{\operatorname{argmin}} \sum_{k=1}^{M} dist(p'_k, g'_{i,k})$$

where $dist(p'_k, g'_{i,k})$ represents the Euclidean distance between p'_k and $g'_{i,k}$.

One notable problem in the matching phase is the large number of outliers (e.g. due to self-occlusion by hairs or some sensing errors like holes and spikes). We handle these outliers by imposing a universal threshold ($t = 50$) to all pixels. If the distance between 2 pixels is larger than the threshold, we regard it as an outlier and thus ignore its information.

3. Experiments

To assess the performance of the proposed system, we built a database using a PrimeSensor and performed a series of experiments on this dataset.

3.1. Data and setup

We collected 1054 frontal faces from 20 people (10 to 135 faces for each person) for testing. Each person is asked to sit in front of a PrimeSensor (0.8m-1.2m) for a short period of time in an office environment, potentially with slight head/facial movements. We randomly selected one face from each person as the gallery face of the enrolled identity. All other faces are tested as probe faces for identification. The canonical faces used for face registration are randomly selected from the gallery. The number of canonical faces in the system can be changed according to different configurations.

3.2. Results and analysis

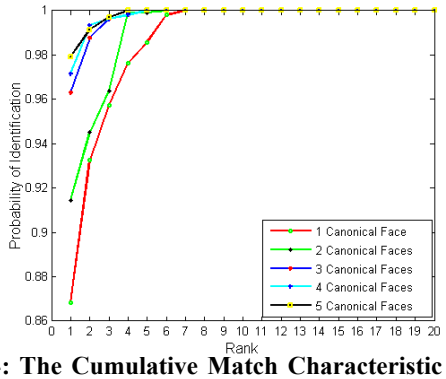

Fig. 4: The Cumulative Match Characteristic (CMC) using 1-5 canonical faces for registration ($M = 1{:}5$).

We first show the Cumulative Match Characteristic (CMC) [16] of the proposed system using various values of M (1 to 5) canonical faces for registration in Fig. 4 (please note that using 1 canonical face is not equal to using AFM or ICS for the registration since the canonical faces were randomly selected). It is clear that the identification rates converge to 100% faster (in fewer ranks) when more canonical faces are incorporated. However, using more than 3 canonical faces does not produce a significant improvement.

Fig. 5 shows the rank-1 identification rates of the proposed system when using *M=1-5* canonical faces for face registration. In the 'Image-to-Image' scenario, each probe face obtained from 1 depth shot is compared with the gallery faces. In the figure, higher identification rates are achieved when more canonical faces are used. When using 5 3D face models as canonical faces, we obtained the highest identification rate in rank-1 (up to 97.91%).

In the 'Video-to-Image' scenario, instead of taking one shot of a person as a probe image, we take a short video sequence of the person as the probe video (9 to 134 frames, which are matched with the 20 gallery faces). First, each frame in the probe video is identified individually; then their results are combined by taking the majority decisions to output the identity of the probe video. In the figure, even if only one canonical face is used, we achieve 100% identification rate. This is because the identification result of each frame is already reliable (as we have shown in the 'Image-to-Image' case), and errors are uniformly distributed for each person (which means that we do not have a lot of errors in one frame sequence but no error for the others). These numbers (100%) are for a 20 persons database. Nevertheless, our results suggest that when extending our system with larger number of subjects in the gallery, the 'Video-to-Image' method would ameliorate the identification degenerations due to increased identity variations.

The proposed system needs to work in real time. Since the system is implemented for a low-cost depth camera (the PrimeSensor), it should perform at interactive rates for daily-use purpose. We conducted our experiments on a consumer-level hardware (Intel(R) Xeon(R) CPU E5520 @ 2.27GHz, NVIDIA(R) Tesla(R) C1060). The average computing time is {0.047s, 0.130s, 0.233s, 0.296s and 0.378s} to identify one probe face when using 1-5 canonical faces for face registration. To test our system performance on a larger dataset, we artificially generated 1000 gallery faces by data duplication, and obtained 1.54s average identification time using 3 canonical faces. To conclude, the proposed face identification system is working in real time.

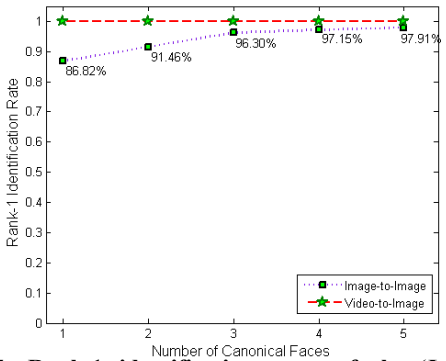

Fig. 5: Rank-1 identification rates of the 'Image-to-Image' approach and the 'Video-to-Image' approach based on 1-5 canonical faces ($M = 1{:}5$) for registration.

According to the results obtained, 3 canonical faces could be the best choice for the proposed face registration method (as well as for the entire system), which balance the identification results and the computing time.

4. Conclusion

We presented a complete framework of an online 3-D face identification system based on PrimeSensor working in real-time. A fast face registration approach using multiple intermediate references is applied and thus the system accuracy is significantly improved. Our system achieves high identification rates on noisy, incomplete and low-resolution face data and can process up to 20 fps on consumer-level hardware.

References

[1] P. J. Phillips, P. J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR), vol.1, no., pp. 947- 954, 20-25 June 2005.

[2] G. Medioni, R. Waupotitsch, "Face Modeling and Recognition in 3-D," in Proc. of AMFG 2003, 232-233, 17 Oct. 2003.

[3] G. Medioni, J. Choi, C.H. Kuo, D. Fidaleo, "Identifying Noncooperative Subjects at a Distance Using Face Images and Inferred Three-Dimensional Face Models," IEEE Tran. on Sys., Man, and Cyber., Part A: Systems and Humans, Vol. 39, No. 1, pp. 12-24, Jan. 2009.

[4] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, "Real-time human pose recognition in parts from single depth images," IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR), vol., no., pp.1297-1304, 20-25 June 2011.

[5] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments," in Proc. of International Symposium on Experimental Robotics (ISER), 2010.

[6] P. J. Besl, N. D. McKay, "A Method for Registration of 3-D Shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, No. 2, pp. 239-256, 1992.

[7] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images," Image Vision Comput., 10(3), 145–155, 1992.

[8] L. Spreeuwers, "Fast and Accurate 3D Face Recognition," Int. J. Comput. Vision, Vol.93, Issue 3, 389-414, July 2011.

[9] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn, "Using multi-instance enrollment to improve performance of 3D face recognition," Computer Vision and Image Understanding, 112(2), 114–125, Nov. 2008.

[10] I. A. Kakadiaris, G. Passalis, G. Toderici, M. N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis, "Three –dimensional face recognition in the presence of facial expressions: an annotated deformable model approach," IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4), 640–649, April 2007.

[11] S. Granger, X. Pennec, "Multi-scale EM-ICP: A Fast and Robust Approach for Surface Registration," in Proc. of 7th European Conference on Computer Vision (ECCV2002), Vol. 4, pp. 69-73, 2002.

[12] T. Tamaki, M. Abe, B. Raytchev, K. Kaneda, "Softassign and EM-ICP on GPU," in Proc. of the 2nd Workshop on Ultra Performance and Dependable Acceleration Systems (UPDAS), 2010.

[13] P. Viola, and M. J. Jones, "Robust Real-Time Face Detection." Int. J. Comput. Vision, Vol. 57, Issue 2, 137-154, May 2004.

[14] nVIDIA, CUDA CUBLAS Library, 2010.

[15] J. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching," Comm. ACM, vol. 18, no. 9, pp. 509-517, 1975.

[16] A. K. Jain and S. Z. Li. Handbook of Face Recognition. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 2005.