Comparing Strategies for 3D Face Recognition from a 3D Sensor*

Jongmoo Choi, Ayush Sharma, and Gérard Medioni[1]

***Abstract*— We address the problem of 3D face recognition from 3D data, using different strategies. One strategy (1F-NF), explored earlier, is to match each individual frame to a set of reference frames. A second one (1F-3D) is to replace the set of reference frames by a 3D model resulting from the integration of individual frames. A third strategy (3D-3D) is to use a 3D face model inferred from multiple frames as the input probe. We show that the recognition performance using 3D model to 3D model outperforms the others, at the cost of a delay in response, due to the model building step.**

I. INTRODUCTION

Face recognition has been an active research topic for several decades and various techniques have been presented [1][2][3][4]. Traditional 2D image-based face recognition methods appear to be sensitive to variations in pose, illumination and expression changes [1]. Since pixel intensity is a non-linear combination of the geometry, viewpoint, lighting, and surface properties, capturing invariant features from projected images is a difficult problem.

Many researchers have presented 3D face recognition methods, as the shape information is independent of viewpoint and lighting changes [25]. Most existing methods use laser scanning [1], stereo vision [2], structure from motion [3][5], or generic face model [8][6][7] to obtain 3D face models. A laser scanner is slow and expensive. Multiple image-based approaches are instable and suffer from costly processing. Recent success of low-cost depth cameras [11], such as PrimeSense camera [27][28], enables to process RGB and depth video stream for 3D face recognition.

We have previously presented a real-time 3D face identification system using a low-cost depth camera in which both an input probe and a set of gallery data are registered with a small number of reference faces in order to reduce computational complexity while preserving recognition rate [9]. We have also presented an accurate 3D face modeling technique that produces a laser scan quality 3D face model from a noisy depth video stream by aggregating registered 3D data into a 2D unwrapped cylindrical coordinate system [10]. Clearly, the performance of 3D recognition system depends on the quality of the input data [25], and accurate 3D face models should enable us to improve the recognition rate. However, the best strategy for combining multiple input frames is not obvious.

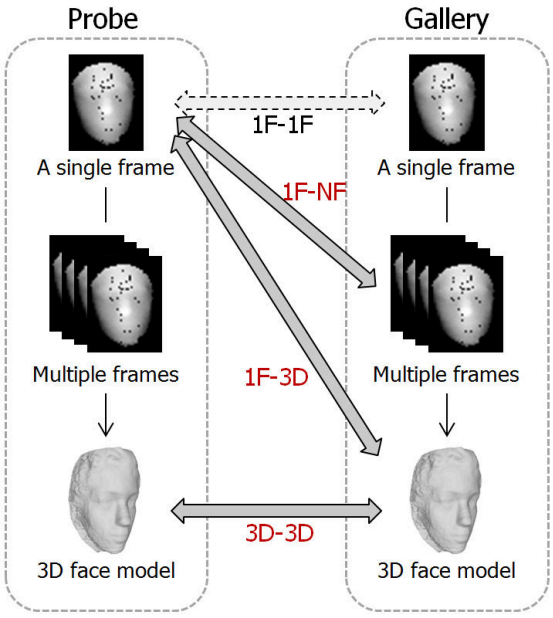

Fig. 1. Three Strategies for 3D face recognition: 1F-NF (single vs. N frames), 1F-3D (single vs. 3D), and 3D-3D (3D face vs. 3D face). We use 1F-1F (single vs. single frame) as a baseline performance.

Of course, one can argue that using more depth data provides better performance because it has more information. It is also shown in our previous work [9]. In contrast, we might lose some information during the modeling process because we use a 2D unwrapped cylindrical system that allows us to represent only star-shape objects [10]. Hence, one important question is whether a system using reconstructed 3D face model performs better than a system using the raw depth frames used for the reconstruction. Our hypothesis is that our modeling should provide better results since it enhances the signal to noise ratio up to a certain point by aggregating multiple observations. To answer to the question, we need a comparison between a method using multiple frames (1F-NF) and a method using a single 3D face mode generated from the same frames (1F-3D).

It is also possible to input multiple frames from an user in many practical applications including human-robot interactions [23]. In this case, we can use all the raw depth images as the probe set (NF-1F) or we can build an accurate 3D face model for the probe (3D-1F or 3D-3D). Because of the symmetric nature of the matching process, the performance of NF-1F strategy can be replaced by the result of 1F-NF.

*This work was partly supported by the IT R&D program of MKE & KEIT [10041610, The development of the recognition technology for user identity, behavior and location that has a performance approaching recognition rates of 99% on 30 people by using perception sensor network in the real environment]

[1]J. Choi, A. Sharma, and G. Medioni are with the Institute for Robotics and Intelligent Systems, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089-0273, USA {jongmooc, ayushsha, medioni} at usc.edu

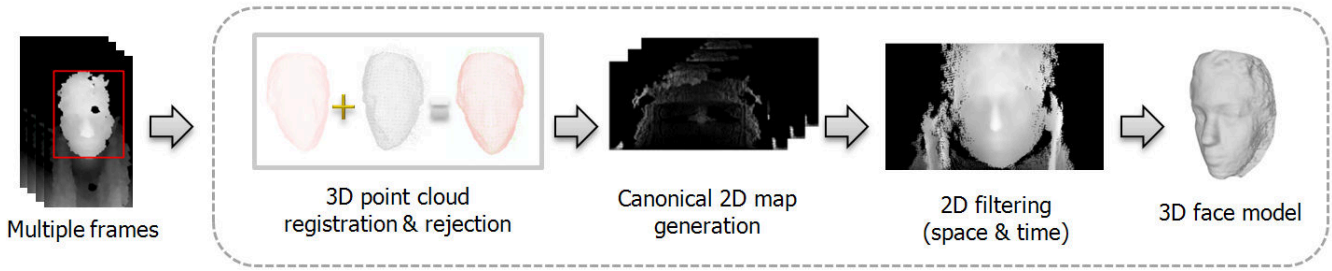

Fig. 2. Overview of the 3D face modeling framework from multiple depth images.

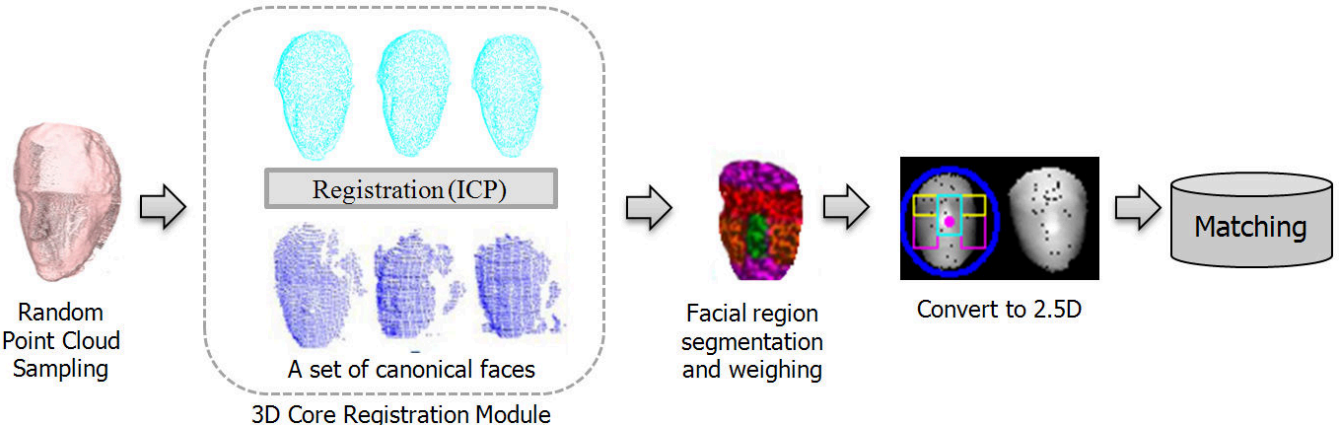

Fig. 3. Overview of the 3D face identification framework using a set of canonical faces.

Similarly, the performance of 3D-1F can be predicted by the 1F-3D.

We provide experimental comparison results between three key strategies: 1F-NF (a single frame vs. multiple frames), 1F-3D (a single frame vs. a 3D face model), and 3D-3D (a 3D face model vs. a 3D face model). The three methods are illustrated in Fig. 1.

We use multiple frames as a gallery data set in the 1F-NF strategy. We use a single 2.5D input probe but match it against a number of gallery images, instead of using a single shot gallery image or a gallery image generated from a 3D face model.

We use a 3D face model as a galley datum in the 1F-3D strategy. We take a noisy sequence of depth frames as input, produce a laser scan quality 3D face model, and generate a gallery image.

In the 3D-3D strategy, the probe is a reconstructed 3D face model produced from a sequence of depth frames. Then, the probe is matched against stored gallery images generated from 3D face models.

In all strategies, we use the same matching framework [9]. To generate a gallery image, we take a 3D (either from a 3D model or a raw depth frame) data, register with a small number of canonical faces, convert to a 2.5D template image, and store it as a gallery datum. Given a probe data, we register it against the same canonical reference faces, and generate a 2.5D pose normalized input face. Then, the normalized probe image is matched against the stored gallery images. The identification framework [9] considerably reduces processing time while preserving the recognition rate if we have a large number of data. In contrast, the standard 3D-3D registration [2][12][13] is very expensive because it requires the nearest neighbor search between thousands of 3D points. For instance, the 1F-NF method needs numerous data for a subject.

Contributions

- We present extensive comparison results between three strategies for 3D face recognition: 1F-3D, 1F-NF, and 3D-3D.
- We present a novel 3D face identification framework, using a real-time depth camera, which builds an accurate 3D face model from a noisy depth stream and uses for the probe and gallery.
- We present validation results on a real-world 10 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. Then, we draw the conclusion in section 4.

II. ALGORITHMIC COMPONENTS

The algorithmic components of the 3D face recognition consists of face detection & segmentation, 3D face modeling from multiple depth images, and identification using a set of canonical faces.

Fig. 4. Reconstructed 3D face models from different subjects.

A. Face detection and segmentation

The output from the PrimeSensor includes a RGB image and a depth map at 640×480 resolution. We use both the RGB image and depth map for face detection. Firstly, we use an implementation of Viola-Jones' method [19][29] using a RGB image. The result might include many false alarms, especially from the background. We check the depth information of each detection result and select the closest face from the sensor. Since we have absolute scale information, we can segment a facial region using a rectangular region from the center of a detected face. The following registration is not sensitive to the error in this segmentation step.

B. 3D face modeling from multiple depth images

The main idea of modeling is to accumulate several depth data containing different poses to compensate for the noisy depth data and refine noisy information through time [10]. An overview of 3D face modeling is presented in Fig. 2. The first frame, containing a near frontal face, is set as a reference and is used to generate a canonical 2D map using a cylindrical representation [5][24]. Each new input depth frame is converted into a 3D point cloud, registered with the reference point cloud, and accumulated into the 2D canonical map. This 2D representation enables to perform 2D image-based operations such as the bilateral filtering to filter out noisy input instead of using complex and expensive 3D mesh processing methods. We detect the nose tip [1] from a reconstructed 3D face, segment the frontal part of the face, and convert to a gallery datum.

[1]We assume that gallery faces are frontal. For a probe face, the nose tip is estimated by the registration result between the input probe and a reference face.

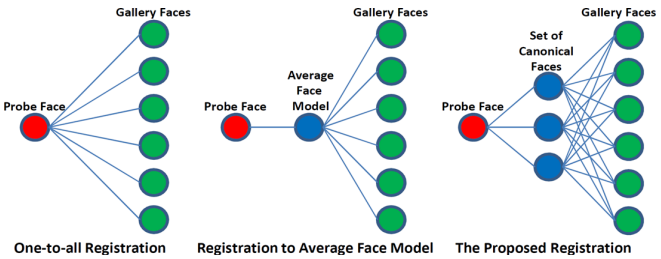

Fig. 5. Architecture of one-to-all registration (left), AFM or ICS (middle), and registration using multiple references (right) [9]. We use the registration using multiple references for all strategies.

1) 3D point cloud registration: The key idea is to use registration between the input point cloud and the point cloud from the reference frame. This method gives bounded error while several registration steps between consecutive images would propagate errors. Although the use of a single reference frame limits the pose angle of the registration, the pose range can be extended using facial halves or multiple reference frames. The rigid transformation is computed by EM-ICP on CUDA [18], which enables to obtain real-time performance. We can use the transformation matrix at time (t-1) as the initials for the registration at time t when we process a video stream. Our system runs at 6 frames per second on a GeForce GTX460. We use around 1,000 points for each frame, which gives a good trade-off between speed and accuracy.

2) Canonical 2D map and filtering: The main idea is to accumulate depth information through time using canonical 2D maps. We use a cylindrical representation to convert a 3D point cloud into a 2D map, as in [5][24]. The geometry of a facial surface is represented using an unwrapped cylindrical depth map. This method enables to aggregate unlimited number of 3D point cloud in a fixed memory and to apply 2D image-based filtering operations. Once we have enough data, the 2D map is transformed to the original coordinate system and we have a 3D face model (See Fig. 4).

3) Converting to a gallery datum: Although our 3D face models cover some parts of hair and the side of the face, for identification, we use only the central part of the 3D surface including the eyes, the nose, and the mouth. Given a 3D face model, we detect the nose tip and segment the central part of the face using a constant sphere (see Fig. 6). The segmented point cloud is stored as a gallery datum.

C. Registration using multiple references

A registration between two 3D point clouds (a probe and a gallery) is essential. To aim real-time 3D face identification, we use an efficient registration framework using a small number of references [9]. An overview of 3D face identification is presented in Fig. 3.

The one-to-all registration method [15] is too computationally expensive for a large dataset. The AFM [16] or ICS method [14], using a single average face, cannot provide

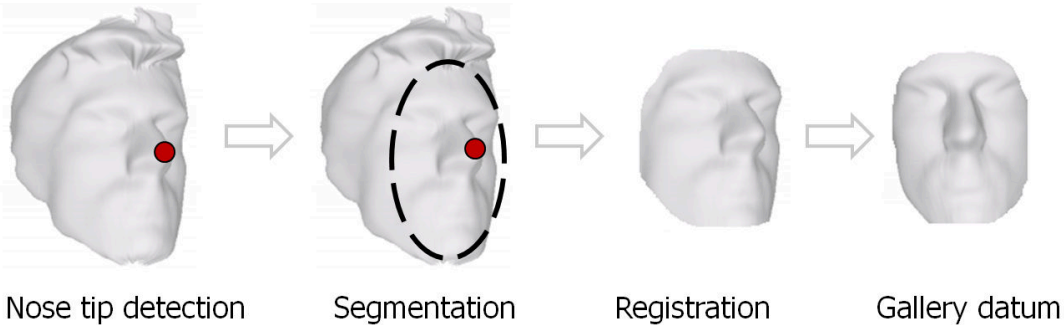

Fig. 6. Illustration of the segmentation of a 3D face model.

accurate registration when we have noisy data from low-cost sensors. First, we randomly select M faces from the gallery to form a set of canonical faces [2]. Each gallery face g_i ($i = 1 : N$, where N is the size of gallery) is aligned with the M canonical faces (using ICP), and thus generates M aligned gallery faces $\{g'_{i,k}, k = 1 : M\}$. A probe face p is also aligned with the same M canonical faces, and thus generates M aligned probes $\{p'_k, k = 1 : M\}$ (See Fig. 5). Then, an aligned probe face is matched with the aligned gallery faces which are aligned with the same canonical faces. This procedure provides better registration than the AFM or ICS method because it is likely that the probe is close to one of the references than the average face and this gives smaller variance of registration. While the registration of all gallery faces (which needs $N \times M$ times alignments) is done during off-line training, an online query requires only M alignments ($M << N$). For real-time performance, we use a GPU version of ICP implementation: EM-ICP algorithm [17][18][20] which runs at $5 \sim 6$ frames per second.

D. Facial region segmentation

We use only the central part of a facial region for identification. After registration, we assume the nose-tip of a probe face is aligned with the canonical face's nose-tip that is manually annotated. Note that this manual annotation is done once in an off-line step. Then the central part is segmented by masking the registered probe face based on a Euclidian distance from the nose-tip.

E. Convert from 3D to 2.5D representation

To reduce the computational complexity of matching two 3D faces, we convert a registered 3D point cloud into a 2.5D representation via orthographic projection. The pixel-wise comparison of two 2.5D images does not require any explicit indexing [21] and thus the matching is much faster than a matching between two 3D point clouds. These 2.5D images registered with the multiple references are sufficient to establish identity, without further feature extractions as shown in [14][16]. We further divide the 2.5D facial images into different facial areas (the nose, eyes region, cheeks region and the rest part) [16]. Each area is associated with a constant weight to indicate its importance in face identification.

[2]We use 4 canonical faces in all experiments.

F. Identification

Given a set of 2.5D probe facial images registered to M canonical faces $\{q_k, k = 1 : M\}$ and N number of 2.5D gallery facial images $\{h_{i,k}, i = 1 : N, k = 1 : M\}$, we find the probe's identity as:

$$id(p) = \arg \min_{i=1:N} \sum_{k=1}^{M} dist(\mathbf{q}'_k, \mathbf{h}'_{i,k}), \quad (1)$$

where $dist(\mathbf{q}'_k, \mathbf{h}'_{i,k})$ represents the Euclidean distance between $\mathbf{q}'_k$ and $\mathbf{h}'_{i,k}$, and $\mathbf{q}'_k$ represents a vectorized image from 2.5D image q_k.

G. Decision from multiple frames

The 3D-3D strategy requires a set of frames to build a 3D face probe. Also, a NF-1F method requires a set of input frames. We can classify these methods as a window-based decision method. In contrast, 1F-1F, 1F-NF and 1F-3D methods use only a single frame as the input. It can be classified as a frame-based decision method. Fig. 7

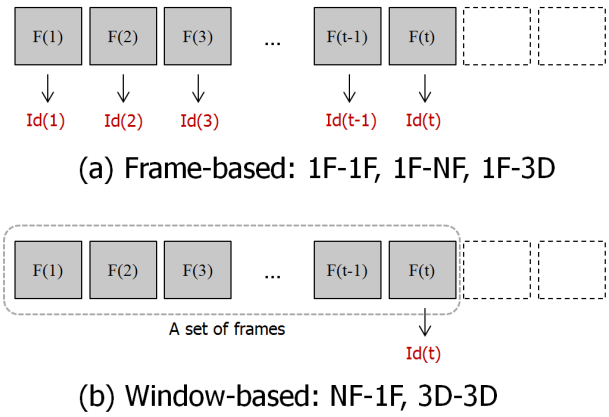

Fig. 7. Single frame-based vs. window-based decision. A window-based method should provide better results than a frame-based one while there might be delayed responses.

illustrate the difference between frame-based and window-based methods.

In general, a window-based method can provide more accurate results than a frame-based one because the neighbor frames are highly correlated and a window-based method utilizes a data fusion technique [26]. Hence, it is expected that both the 1F-NF and the 3D-3D should provide better results than a 1F-1F method while the window-based methods might have significantly delayed responses.

III. Experiments

We present a comparison result between three strategies for 3D recognition: using a single frame vs. multiple frames (1F-NF), a single frame vs. a 3D face model (1F-3D), and a 3D face model vs. a 3D face model (3D-3D).

A. Data and setup

We collected 20 RGB+D videos from 10 people (two videos per subject) for testing. Each person is asked to sit in front of a Primesense camera $(0.8m - 1.2m)$ for a short period of time ($15 \sim 20$ seconds) in an office environment with wide out-of-plane head movements (yaw angle: $\pm 60^{\circ}$). Each video clip contains around $200 \sim 250$ depth frames.

We built three gallery datasets from all the first videos. The first gallery database includes 10 reconstructed 3D face models from the subjects:

$$G_{3D} = \{d_1, d_2, \cdots, d_N\},$$

N is the number of subjects. Each 3D face model is segmented based on automatically detected nose tip. The second gallery database consists of a set of depth frames extracted from the first videos:

$$G_{video} = \{g_1^1, g_1^2, \cdots, g_1^k, g_2^1, g_2^2, \cdots, g_N^k\},$$

k is the number of depth frames ($k = 10$) for each subject. From each video, we selected 5 frontal faces and 5 near-frontal faces with $5^{\circ} \sim 10^{\circ}$. The third gallery database contains only a single frontal face for each subject:

$$G_{1F} = \{g^1, g^2, \cdots, g^N\}.$$

We built two probe datasets from all the second videos. The first probe database contains $m(= 5)$ randomly selected frames for each subject from the second videos:

$$P_{1F} = \{p_1^1, p_1^2, \cdots, p_1^m, p_2^1, p_2^2, \cdots, p_N^{m-1}, p_N^m\}.$$

The second probe database consists of reconstructed 3D face models:

$$P_{3D} = \{p_1^3, p_2^3, \cdots, p_N^3\},$$

N is the number of subjects.

The canonical faces used for face registration are randomly selected from the different face database. The number of canonical faces is 4 in all experiments.

B. 1F-1F: single frame vs. single frame

As a baseline performance, we used a single frame both for probe and gallery[3]. We used P_{1F} and G_{video} for probe and gallery set, respectively. We computed the distances between all the pairs of the probe and gallery set:

$$P_{1F} \times G_{video} = \{(p_1^1, g_1^1), (p_1^1, g_2^1), \cdots, (p_N^m, g_N^k)\}$$

and computed statistics independently. We first show the ROC graph of the 1F-1F method in Fig. 8. The result shows 65.1% recognition rate at FAR (False Acceptance Rate) 15%.

C. 1F-3D: single frame vs. 3D face model

The first strategy we want to investigate is using reconstructed 3D faces as a gallery set. The experimental protocol is the same as 1F-1F experiment except that we used reconstructed 3D face models G_{3D} for the gallery set G_{video}. The result shows 55.6% recognition rate at FAR 15%.

D. 1F-NF: single frame vs. multiple depth frames

In this 'Image-to-Video' scenario, instead of using one shot of a person as a gallery image, we can take a short video sequence of the person as the gallery video. In this experiment, we selected 10 frames per subject. First, each frame in the gallery video is identified individually; then their results are combined by taking the minimum distance to output the identity of the gallery video as:

$$dist(p, g_j) = \min\{dist(p, g_j^1), dist(p, g_j^2), \cdots, dist(p, g_j^k)\}, \quad (2)$$

where k is the number of frames for j-th gallery data. The result shows 86.4% recognition rate at FAR 15%. 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 $0.296s$ to identify one probe face when using 4 canonical faces for face registration. Hence, we can extend our system with a large number of subjects in the gallery.

E. 3D-3D: 3D face model vs. 3D face model

Finally, we show the recognition performance using 3D face models for both probe and gallery. We used P_{3D} and G_{3D} for probe and gallery, respectively. The experimental results show that both 1F-NF and 3D-3D outperforms the baseline method (1F-1F) across all threshold values and the 3D-3D method is slightly better than the 1F-NF method. The 3D-3D shows 90.0% recognition rate at FAR 15% while the 1F-NF shows 86.4% at the same FAR rate. Each 3D face model contains around $6.5K$ 3D points that require $181KB$ memory space while each gallery in 1F-NF method needs about $650KB(65KB \times 10 frames)$ memory.

[3]The 3D face identification software is available from [30].

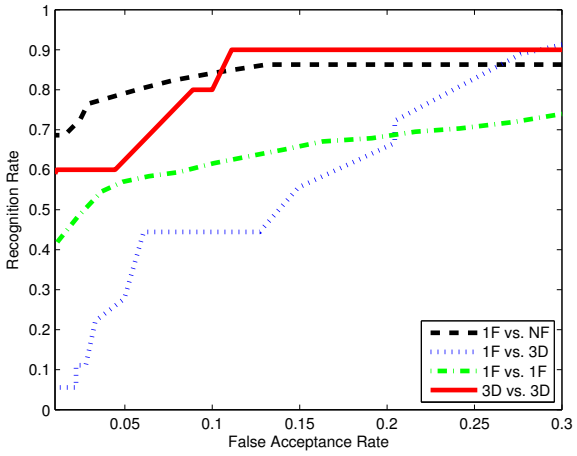

Fig. 8. Comparison results.

F. Processing time and response time

A computation time of an implemented 3D recognition system can be explained by processing time and response time. Given a 1F input, the probe image generation takes 0.67 seconds with 4 reference faces. In case of 3D probe, it takes 3.63 seconds with the same number of 3D face references, without any optimization. The processing time of 3D probe can be significantly reduced if we use a simple data sampling method and a GPU implementation. For identification, both 1F probe and 3D probe take only 0.01 seconds because the same size of 2.5D images are compared in the process.

We can build a 3D face model from a live video stream even though each 3D probe was taken from a video ($15 \sim 20$sec) in the experiments. Building a 3D face model from a live depth stream would take $4 \sim 5$sec. However we cannot avoid this response time in the 3D-3D method, due to the model building step, while a single 1F probe can be captured in $60ms$ $(15Hz)$.

IV. CONCLUSION

We present comparison results between three strategies for 3D face recognition from a real-time, low-cost 3D sensor. The 1F-NF strategy (matching each individual frame to a set of reference frames), the 1F-3D strategy (replacing the set of reference frames by a 3D model), and the 3D-3D strategy (using reconstructed 3D face models as the input probe and gallery) surpass the baseline 1F-1F method (using a single probe and a single gallery). It is confirmed that the recognition performance using 3D model to 3D model outperforms the others, at the cost of a significant delay in response, due to the model building step.

REFERENCES

[1] P. J. Phillips, P. J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Wore, Overview of the face recognition grand challenge, IEEE CVPR, 2005.

[2] G. Medioni and R. Waupotitsch, Face Modeling and Recognition in 3-D, AMFG, 2003.

[3] J-M. Choi, C-H Kuo, D. Fidaleo, and G. Medioni, Identifying Non-cooperative Subjects at a Distance Using Face Images and Inferred Three-Dimensional Face Models, IEEE Transactions on Systems, Man, and Cybernetics(SMC)-A, Vol. 39, No. 1, pp. 12–24, January 2009.

[4] Hoda Mohammadzade and Dimitrios Hatzinakos, Iterative Closet Normal Point for 3D face Recognition, IEEE TPAMI, Vol. 35, No. 2, 2013.

[5] Yuping Lin, Grard Medioni, and Jongmoo Choi, Accurate 3D Face Reconstruction from Weakly Calibrated Wide Baseline Images with Profile Contours, IEEE CVPR, 2010.

[6] Jongmoo Choi, Gerard Medioni, Yuping Lin, Luciano Silva, Olga Regina Pereira Bellon, Mauricio Pamplona Segundo, and Timothy Faltemier, 3D Face Reconstruction Using A Single or Multiple Views, ICPR, 2010.

[7] Jongmoo Choi, Yann Dumortier, Sang-Il Choi, Muhammad Bilal Ahmad, and Gerard Medioni, Real-time 3-D Face Tracking and Modeling From a Webcam, IEEE WACV, 2012.

[8] Volker Blanz and Thomas Vetter, Face Recognition Based on Fitting a 3D Morphable Model, IEEE TPAMI, PP. 1063–1074, 2003.

[9] Rui Min, Jongmoo Choi, Gerard Medioni, and Jean-Luc Dugelay, Real-Time 3D Face Identification from a Depth Camera, ICPR, 2012.

[10] Matthias Hernandez, Jongmoo Choi, and Gerard Medioni, Laser Scan Quality 3-D Face Modeling Using a Low-Cost Depth Camera, EUSIPCO, 2012.

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

[12] P. J. Besl and N. D. McKay, A Method for Registration of 3-D Shapes, IEEE TPAMI, Vol. 14, No. 2, pp. 239–256, 1992.

[13] Y. Chen and G. Medioni, Object modelling by registration of multiple range images, Image Vision Comput., Vol. 10, No. 3, pp. 145–155, 1992.

[14] L. Spreeuwers, Fast and Accurate 3D Face Recognition, Int. J. Comput. Vision 93, pp. 389–414, 2011.

[15] 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, vol. 112, No. 2, pp. 114–125, 2008.

[16] 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 TPAMI, Vol. 29, No. 4, pp. 640–649, 2007.

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

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

[19] P. Viola, and M. J. Jones, Robust Real-Time Face Detection, Int. J. Comput. Vision Vol. 57, No. 2, pp. 137–154, 2004.

[20] nVIDIA, CUDA CUBLAS Library, 2010.

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

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

[23] K. Fukui and O. Yamaguchi, Face recognition using multi-viewpoint patterns for robot vision, Robotics Research, pp. 192–201, 2005.

[24] L. Williams, Performance-driven facial animation, Computer Graphics, vol. 24, no. 4, 1990.

[25] Kevin W. Bowyer, Kyong Chang, and Patrick Flynn, A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition, Computer Vision and Image Understanding, Vol. 101, No. 1, pp. 1–15, 2006.

[26] A. Hadid and M. Pietikainen, From still image to video-based face recognition: an experimental analysis, Proceedings. Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 813–818, 2004.

[27] http://www.primesense.com/.

[28] http://www.openni.org/.

[29] http://opencv.org/.

[30] http://www.openni.org/files/3d-face-identification/, OpenNI2 Middleware Libraries- 3D Face Identification by University of Southern California, 2012.