Lock3DFace: A Large-Scale Database of Low-Cost Kinect 3D Faces

Jinjin Zhang *Di Huang* * *Yunhong Wang* *Jia Sun*

IRIP Lab, School of Computer Science and Engineering,
Beihang University, Beijing, 100191, China
{zhang_jhon,dhuang,yhwang,sunjia}@buaa.edu.cn

Abstract

In this paper, we present a large-scale database consisting of low cost Kinect 3D face videos, namely Lock3DFace, for 3D face analysis, particularly for 3D Face Recognition (FR). To the best of our knowledge, Lock3DFace is currently the largest low cost 3D face database for public academic use. The 3D samples are highly noisy and contain a diversity of variations in expression, pose, occlusion, time lapse, and their corresponding texture and near infrared channels have changes in lighting condition and radiation intensity, allowing for evaluating FR methods in complex situations. Furthermore, based on Lock3DFace, we design the standard experimental protocol for low-cost 3D FR, and give the baseline performance of individual subsets belonging to different scenarios for fair comparison in the future.

1. Introduction

Face Recognition (FR) is one of the most active topics in the domain of pattern recognition and computer vision for a number of scientific challenges and a wide range of industrial applications. Compared to other biometric traits, such as fingerprint and iris, face is less intrusive and more natural for identifying a person. In the past several decades, 2D image based FR has achieved tremendous progress [31, 32]; however, it is still not reliable enough because of the unstable performance in complicated environment especially when illumination and pose variations occur. With the rapid development of 3D imaging systems, 2.5D or 3D scans have emerged as a major alternative to deal with the issues that are unsolved in 2D FR, since they deliver geometric cues of faces. In the last ten years, 3D FR has been extensively discussed, with its accuracies of public benchmarks being greatly boosted. Many studies reveal that 3D FR not only reaches competitive results itself [13, 20, 21, 22, 35], but also shows good complementarity to that in the 2D modality [14, 23]. See surveys [7, 16, 30] for more details.

The models used in the state of the art 3D FR systems are of high-quality as the ones in FRGC [28], Bosphorus [29], BU-3DFE [37], *etc.*, recorded by sophisticated equipments. In the early years, the devices to capture such data, *e.g.* Minolta VIVID 910, may take a few seconds for a single session, and during this period, faces are required to keep still, which makes it unsuitable for on-line FR scenarios, especially when users are not so cooperative. Along with the continuous development in both hardware and software, the following versions, *e.g.* 3dMD and Artec3D, are able to provide dynamic flows of 3D face scans of a high resolution at the rate of tens of frames per second. But they are at rather high prices, generally hundreds or even thousands of times more expensive than 2D cameras. Moreover, they are usually big in size and not convenient to operate and it thus leaves a hard problem to implement systems based on them in practical conditions.

The recent advent of low-cost and real-time 3D scanning devices, such as Microsoft Kinect and Asus Xtion Pro Live, makes it possible to collect and exploit 3D data in our daily life, where the sampling precision is decreased to balance the price. Due to the popularization of Kinect sensors, low cost 3D data (or with the texture counterpart, *i.e.* RGBD data) have received increasing attention in the academia in various aspects, including action recognition, object detection, scene classification, *etc.* In contrast to the aforementioned tasks, FR using low cost 3D data is more challenging, because the compromise between cost and accuracy in Kinect makes the data much more noisy, leading to serious loss of discriminative details. Some preliminary attempts on such an issue have been made, and the best result is up to 100% [24], indicating its feasibility to some extent. Nevertheless, the score is not sufficiently convincing, because the subjects in the evaluation dataset are not many enough and only with limited variations.

In this paper, we present a large-scale dataset of low cost Kinect faces, namely Lock3DFace, aiming to thoroughly investigate low cost 3D FR and comprehensively compare

[*] Corresponding Author

Table 1. Overview of major 3D face databases captured using different scanners.

Name	Sub.	Var.	Size	Device	Device Price	Accuracy(mm)
CASIA (2004) [1]	123	E, P	4624	Vivid 910	>$45k	∼0.05
FRGC (2005) [28]	466	E, Ti	4007	Vivid 910	>$45k	∼0.05
BU-3DFE (2006) [37]	100	E	2500	3dMD	>$50k	<0.2
Bosphorus (2008) [29]	105	E, P, O	4666	Inspeck 3D Mega II	>$11k	∼0.7
BU-4DFE (2008) [36]	101	E	60600^{v}	3dMD	>$50k	<0.2
3D-TEC (2011) [34]	214	E	428	Vivid 910	>$45k	∼0.05
UMB-DB (2011) [8]	143	E, O	1473	Vivid 900	>$45k	<0.2
Florence Superface DB (2012) [2]	50	P	50^{v}	Kinect	$249	2-4
KinectFaceDB (EURECOM) (2013) [17]	52	E, I, P, O, Ti	936	Kinect		
3D Mask Attack DB (2013) [10]	17	Ti	255^{v}	Kinect		
CurtinFaces (2013) [19]	52	E, I, P, O	4784	Kinect		
FaceWarehouse (2014) [5]	150	E	3000	Kinect		
Lock3DFace (2015)	509	E, I, P, O, Ti	5711^{v}	Kinect V2	$199	∼2

* E: Expression, I: Illumination, P: Pose, O: Occlusion, Ti: Time, ^{v}: Video clips are recorded while the others capture only single frames.

the approaches. To the best of our knowledge, Lock3DFace is the largest database of low-cost 3D face models publicly available, which consists of 5711 video samples with a diversity of variations in expression, pose, occlusion, and time lapse, belonging to 509 individuals. In each raw face record, the clues in the texture and near infrared modalities are also provided, supporting the scenarios of 2D FR, near infrared FR, multi-modal FR, and heterogeneous FR.

Besides the Lock3DFace database, there are two additional contributions in this paper:

1) We manually label some fiducial landmarks, *i.e.*, the nose tip, the corners of eyes, and the inner corners of mouth, if there are visible, on the first frame of each video clip of 3D faces, and present a preprocessing pipeline to deal with the noisy models;

2) We design a standard experimental protocol for low cost 3D FR, and propose a method making use of Iterative Closet Points (ICP) to match better quality face models reconstructed from low cost video clips, based on which baseline results are given.

The remainder of the paper is organized as follows. We briefly review the existing low cost 3D face databases and highlight the properties of Lock3DFace in Sec.2. In Sec.3, we introduce the Lock3DFace database in detail, including acquisition, challenges, protocols, and preprocessing. Sec.4 presents the method of reconstructed model based 3D FR. Baseline experimental results are displayed and analyzed in Sec.5. Finally Sec.6 concludes the paper with perspectives.

2. Related Work

Table 1 summarizes the major 3D face databases, including the information of name, complete year, subject number, sample number, challenge, and device with its price and accuracy. As shown in the table, most existing databases are collected by expensive 3D scanners, *e.g.* FRGC is captured by the Minolta Vivid 910 Scanner and BU-3DFE is acquired by the 3dMD imaging system. In recent years, a number of 3D face databases of low cost data have become available. This section gives an up-to-date review of these databases as well as corresponding FR methods on them.

The Florence Superface Dataset [2] comprises a pair of low-resolution and high-resolution 3D face scans of an individual, captured by Kinect and 3dMD respectively, aiming to investigate FR across resolutions, where only 20 subjects are used. Later, they release the version 2.0 of the database [3] that includes 50 subjects. In that work, a super-resolution based reconstruction approach is proposed to build a face model of higher resolution using a sequence of low-resolution 3D face video whose length is 10 to 15 seconds, and the reconstructed model is further matched with the ones given by 3dMD. The results support the idea that constructing super-resolved models from consumer depth cameras can be used in real application contexts, but it is not well demonstrated whether it works in 3D FR due to data limitation.

KinectFaceDB [25] consists of 936 multi-modal facial images of 52 people sensed through Kinect at two different periods (with the interval of about half month). In each session, the dataset provides each person with 9 samples of various expressions as well as lighting and occlusion conditions. Evaluations are conducted using some famous FR methods, including Principal Component Analysis (PCA), Local Binary Patterns (LBP), Scale Invariant Feature Transform (SIFT), Local Gabor Binary Patterns (LGBP), ICP, and Thin Plate Spline (TPS), and the gain in performance is demonstrated when integrating the depth data with the RGB data via score level fusion. Moreover, quantitative comparison of the proposed KinectFaceDB and FRGC is provided, which reveals the imperative needs of such data for FR. Unfortunately, similar to the one above, the size of

KinectFaceDB does not well support further investigation on this issue.

The 3D Mask Attack Database (3DMAD) [10] is built to evaluate face based real-access and spoofing attacks, which contains 255 video clips of 17 persons recorded by Kinect. The data are collected in 3 different sessions for all subjects, and in each session, 5 videos of 300 frames are captured under controlled conditions, with frontal-view of the neutral expression. The authors analyze LBP based approaches both in the color and depth modalities, and the experimental results suggest that for both data types, accurate classification rate can be achieved using Linear Discriminant Analysis (LDA). However, this database is not suitable for FR, since the subjects in it are really few.

The CurtinFaces database [19] includes 4784 facial images of 52 subjects, it is established to validate FR methods in the presence of variations in pose, illumination, facial expressions, and sunglasses disguise. A multi-modal Sparse Representation Classifier (SRC) is proposed for 2D+3D FR, and the proposed system reports satisfactory accuracies under the challenges above. The main drawback lies in that the proposed method requires a large number of gallery faces captured at different times for each individual, which is not always available in the real condition. Furthermore, despite the sharp increase in samples, the number of subjects is still small, and all the samples are collected in one session, making the results are not that reliable.

FaceWarehouse [5] is a database of 3D facial expressions for visual computing applications, and it is composed of 3000 facial images of 150 individuals aged from 7 to 80 of various ethnic backgrounds with the neutral expression and 19 other actions, such as mouth-opening, smile, kiss, *etc*. A template facial mesh is deformed to fit the depth data as closely as possible while matching the feature points on the color image to the corresponding ones on the mesh, and a set of individual-specific facial expression blendshapes is subsequently constructed for each person. Nevertheless, this dataset only considers expression variations, which is far from sufficient for FR in complex situations.

Different from the current 3D face databases acquired by Kinect, Lock3DFace is collected using Kinect II of updated hardware setup. Meanwhile, it has a significant increment in the amount of individuals, and considers all the major challenges in FR. Furthermore, the time lapse between the two sessions is up to 7 months. All these factors make Lock3DFace greatly superior to the ones mentioned above for low cost 3D (or RGBD) FR.

3. Lock3DFace Database

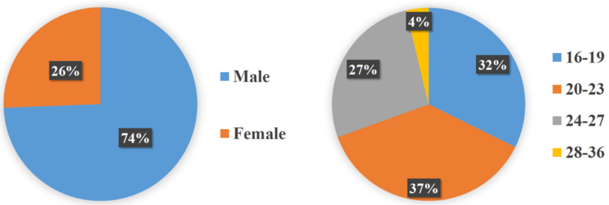

Figure 1. Data distribution of Lock3DFace of (a) gender and (b) age.

The Lock3DFace[*] database totally consists of 5711 RGBD face video clips belonging to 509 individuals with diverse changes in facial expression, pose, occlusion and time lapse. It is mainly designed for low cost 3D FR, but also for 2D FR, near infrared FR, multi-modal (RGBD) FR, and heterogeneous FR. In the subsequent, we introduce its details including acquisition, challenges, protocols, and preprocessing.

3.1. Data Acquisition

Lock3DFace is acquired using Kinect V2, the second-generation of the Kinect sensor. As with the original Kinect, the sensor uses infrared to read its environment, but presents greater accuracy over its predecessor. Kinect V2 updates the 2D camera to a higher resolution one that can be used for color video recording. Moreover, it has an increased field of view, thus reducing the amount of distance needed between the user and the sensor for optimal configuration.

All the data are captured under a moderately controlled indoor environment with natural light in the daytime. The participants are asked to sit in front of the Kinect sensor fixed on the holder and are not allowed to move rapidly when the video is recording for 2-3 seconds. Three types of modalities, *i.e.*, color, depth, and infrared are collected in individual channels at the same time. The color frames are recorded with the size of 1920×1080, and the depth and infrared frames are of the resolution of 512×424. There are in total 509 volunteers who participate in the collection process. Among them, 377 are male and 122 are female. Since the majority of them are undergraduate, master, and Ph.D. students in the universities, their ages distribute in the range of 16 to 36 years old. See Fig.1 for more details. All the major challenges in FR are considered, involving the changes in expression, pose, and occlusion. The dataset contains two separate sessions with a long interval up to 7 months. All the 509 subjects join the first session, while 169 join the second session, thereby presenting time lapse variations as well. Regarding an individual subject, in each session, at least two video clips are made in the categories of neutral-frontal, expression, pose, and occlusion. The details of these challenges are described in Section 3.2.

[*] The database is available on request at http://irip.buaa.edu.cn/.

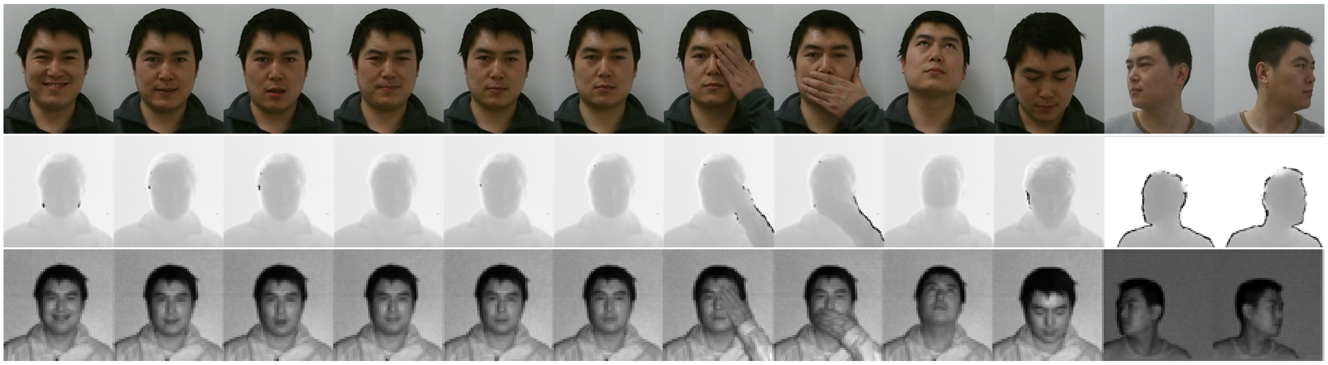

Figure 2. Sample illustration of different challenges in the database. From left to right, happiness, anger, surprise, disgust, fear, neutral face, left face occluded by hand, mouth occluded by hand, looking-up, looking-down, left face profile, and right face profile are displayed in order, where the last two are captured in the second session and the others in the first one. Upper row: RGB images; middle row: depth images; and bottom row: near infrared images which share the same coordinate with the depth maps.

3.2. Challenges

To comprehensively evaluate FR methods, especially to simulate complex conditions in the real world, volunteers are required to present different expressions, poses, and occlusions in each session, forming five categories of frontal-neutral, expression, pose, occlusion, and time. The five parts are introduced in detail respectively in the following.

Neutral-frontal:

The volunteers are scanned in the frontal pose without any expression and occlusion.

Expression:

Six types of universal expressions are considered, including happiness, anger, sadness, surprise, fear, and disgust. The participants are asked to randomly perform at least two kinds of expressions in the frontal pose.

Pose:

The volunteers are required to move their heads in both pitch and yaw directions to an angle up to 90 degrees, displaying the samples with pose variations, where expressions are neutral. The severe ones often lead to self-occlusion with large area missing.

Occlusion:

Only external occlusions are presented where the faces are in the frontal view. People randomly cover up a part of their faces, *e.g.* eye, chin, cheek, or forehead, by hands or wearing glasses.

Time-lapse:

Around a third of all the individuals participate in the second session after 7 months using almost the same configuration as the first one. The previous challenges are also considered, and the distance between the volunteers and the senor is varied to simulate more difficult conditions.

Some examples of a subject are demonstrated in Fig.2, from which we can see that a large diversity of variations are included, and it is a distinct property of Lock3DFace. Table 2 shows its data organization in terms of different variations. Actually, Lock3DFace also has the changes in lighting conditions and radiation intensity in the RGB and infrared channel respectively, because some models of the same subject are recorded at different sessions. But we mainly concentrate on low cost 3D FR, and thus do not highlight this challenge.

Table 2. Data organization of the Lock3DFace database in terms of different challenges.

Variations	Session-1		Session-2	
	Sub.	Sample	Sub.	Sample
Neutral-frontal	509	1014	169	338
Expression		1287		338
Pose		1014		338
Occlusion		1004		338
Total		4319		1352

3.3. Scenarios and Protocols

Despite the Kinect sensor delivers multi-channel images (*i.e.* RGBD), we mainly focus on the depth modality and design the standard experimental protocol for the scenario of low cost 3D FR. The ones of other scenarios, *e.g.* RGB FR, near infrared FR, multi-modal FR, and heterogeneous FR, can be figured out referring to this. To comprehensively evaluate the performance in 3D FR based on Kinect data with respect to various challenges, we group all the samples into several subsets which are described in Table 3. The uniform Gallery_Set includes the first neutral face of each person in Session-1 (S-1). Probe_Set contains four subsets: Probe_Set_1 for all the 1287 face samples with expression changes in S-1; Probe_Set_2 for all the 1014 face samples with pose variations in S-1, Probe_Set_3 for all the 1004 face samples with occlusions in S-1 and Probe_Set_4 for all

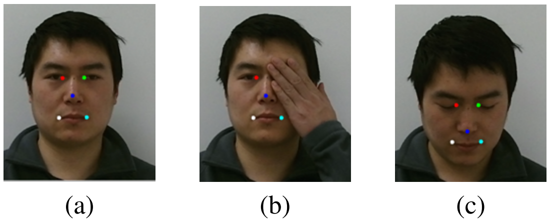

Figure 3. Manually labeled fiducial landmarks, *i.e.* nose tip, inner corners of eyes, corners of mouth, are given if they are visible on (a) a frontal sample; (b) a sample with occlusion; and (c) a looking down sample.

the 1352 faces in Session-2 (S-2).

Table 3. The standard protocol of Kinect based 3D FR for different scenarios on the Lock3DFace database.

Data	Variations
Gallery_Set	$N_1^* \times S_1$
Probe_Set_1	$E \times S_1$
Probe_Set_2	$P \times S_1$
Probe_Set_3	$O \times S_1$
Probe_Set_4	$\{N, E, P, O\} \times S_2$

* E: Expression, P: Pose, O: Occlusion, N: Neutral-frontal, N_1^*: only the first neutral-frontal face video clips included, S_1:Session1, S_2: Session2.

3.4. Preprocessing

To improve the convenience of the researchers to work with Lock3DFace, along with the database, we provide a preprocessed version of the data. On the one hand, some fiducial points are manually marked on the first frame of each RGB and infrared facial video clip respectively, and the corresponding ones on the depth map are then easily obtained due to the point-to-point correspondence with the infrared map. These landmarks are a few distinct anthropometric points shared by all human beings, including the nose tip, two inner corners of eyes, and two corners of mouth, as shown in Fig.3. These points do not always exist for pose variations and external occlusions, and they are labeled only if they are visible. They offer the simplicity in face cropping, pose correction, feature extraction, *etc.* in face analysis. Additionally, such points can be regarded as ground-truth to evaluate the techniques of 3D facial landmarking on low cost data.

On the other hand, the depth images captured by Kinect are very noisy (as the first column in Fig.5 shows), and unlike the RGB data, they cannot be directly used for feature extraction. In this study, we present a pipeline to deal with the low cost 3D data, including spike and outlier removing, hole filling, and smoothing. Specifically, the phase-space method in [26] is employed to exclude the spike. The values of some pixels on the depth map are sensed as 0 when they cannot be precisely measured, *e.g.*, the ones around the margin of the object, even though the true depth values are not. To solve this problem, thresholding is applied, and a non-negative threshold is set in order to remove those unmeasurable pixels, and the missing data can then be filled using the cubic interpolation technique. To remove the noise, we adopt the bilateral filter [33], a simple, non-iterative method which has the property of edge-preserving, and during smoothing, it retains the shape information as much as possible that is supposed to contribute in FR.

4. Baseline Method

As we state in Sec.1, compared with the high-quality 3D face models, the ones captured by the Kinect sensor are of relatively low-resolution and highly noisy. This forms the special challenge in low cost 3D FR. A number of approaches are proposed to produce super-resolution depth images from a single input of low resolution, but they do not make much sense in FR, since the points recovered lack for discrimination. Although some studies [19] show that single low cost 3D face frame based methods are able to achieve good performance, they require a big gallery set and the samples of each subject enrolled are expected to include different variations, which is not always fulfilled in the real condition. Fortunately, Kinect records real-time 3D videos, and a short video clip of several seconds (as the ones in Lock3DFace) can be regarded as a sample in FR. More importantly, a 3D face model of much better quality can be reconstructed from several consecutive frames by 3D accumulation and refining techniques [12, 18]. The super-resolution models can then be directly used to compute the similarity of faces using current 3D FR methods.

Based on this consideration, we propose a baseline approach to 3D FR on the Lock3DFace database, which first reconstructs a better model from the low cost 3D face video and then applies ICP to match the face models for similarity measurement. In [27], the KinectFusion system is introduced, fusing all the depth data streamed from a Kinect sensor into a single global implicit high-quality surface model of the observed scene in real-time. In our case, we only target on the face region rather than the whole scene so that the reconstructed model is optimal for FR.

The volumetric non-parametric surface representation method [9] is used to combine the depth maps of cropped faces. Firstly, the 6 Degrees-of-Freedom (6-DOF) pose of the Kinect sensor is tracked based on the consecutive depth faces by ICP, and in each frame, the depth face map can be transformed into a unique global coordinate space. The Signed Distance Function (SDF) is then exploited to integrate the global 3D vertices on a given frame into the voxels which are pre-defined to represent a fixed 3D physical

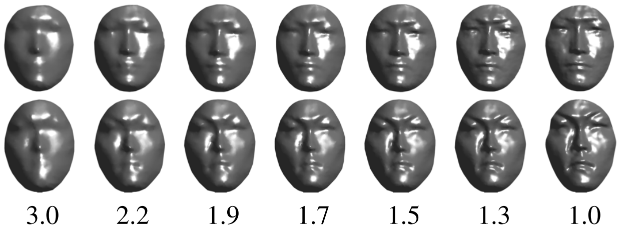

Figure 4. The face models reconstructed based on different resolution that increases from left to right and the numbers below indicate the volumetric size. The larger the number, the lower the resolution.

space. SDF describes the surface as zero and specifies a relative distance to the actual one. Different distance values are stored in the voxels with the positive ones corresponding to the free space and the negative ones to the occupied space. In practice, the Truncated SDF (TSDF) is adopted with the region only around the actual surface stored in the voxels for the sake of efficiency. With respect to each volume slice on z-axis, the current weight $W_k(x, y)$ and truncated signed distance value $D_k(x, y)$ at location (x, y) in the k_{th} frame are updated in parallel as depicted in the formulations below:

$$D_k(x,y) = \frac{W_{k-1}(x,y)D_{k-1}(x,y)+w_k(x,y)d_k(x,y)}{W_{k-1}(x,y)+w_k(x,y)} \tag{1}$$

$$W_k(x,y) = \min(W_{k-1}(x,y) + w_k(x,y), t) \tag{2}$$

$$d_k(x,y) = \begin{cases} \min(1, \frac{s_k(x,y)}{m}) \cdot \mathrm{sgn}(m) & \text{if } s_k(x,y) \geq -m \\ null & \text{otherwise} \end{cases} \tag{3}$$

where $s_k(x, y)$, m are the signed distance and the maximum positive truncated distance respectively; t is the threshold of the max weight; $d_k(x, y)$ is the global frame projective TSDF and $w_k(x, y)$ is its weight set at 1. In addition, the volume size is also important, and it decides the resolution of the reconstructed face models. In this study, we empirically choose the volume size at 1 and the reconstructed precision at 1024.

As illustrated in Fig.4, the reconstructed face model becomes better in the procedure of KinectFusion with higher resolution. Fig.5 demonstrates some face models of an individual reconstructed from the low cost 3D face video clips of different number of frames.

When the 3D face models of high-quality are generated from the low cost video clips, they are regarded as individual samples in the traditional scenario of 3D FR. As in many existing 3D face databases, *i.e.*, FRGC, the baseline results are usually given by the ICP algorithm. To better evaluate the difference between Lock3DFace and the ones captured by expensive 3D scanners, we follow this way, and apply the standard ICP [4] to measure the similarity between different face models reconstructed by KinectFusion. Other versions of ICP that claim to be more competent at 3D FR, such as Region ICP [15], can alternatively be investigated.

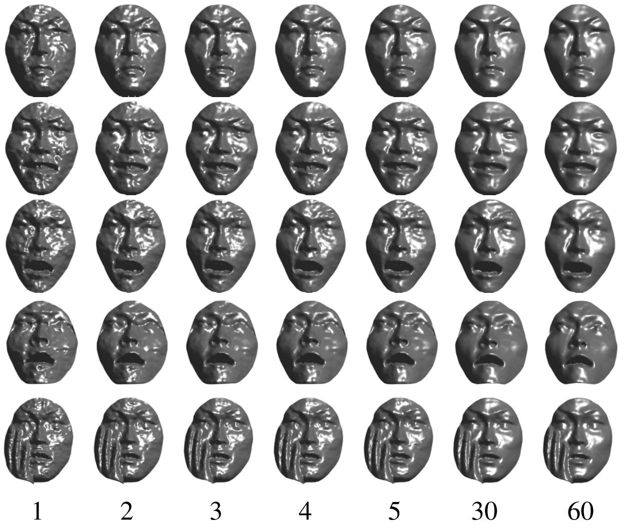

Figure 5. The 3D face models reconstructed as the number of frames increases.

5. Experimental Results

As introduced in Sec.4, a high-quality model is produced from each low cost 3D face video clip using KinectFusion, which is further used to calculate the ICP based registration error for identification. We launch this approach under the standard protocol (presented in Section 3.3) to deliver the baseline results of 3D FR on the Lock3DFace database. The first 60 frames of each 3D video clip are used in KinectFusion to generate the high-resolution 3D face model. Recall that for each individual, we take the first neutral-frontal face model as the gallery sample, and the other models with different types of variations *i.e.* expression, pose, occlusion and time lapse, form four probe sets.

Table 5 displays the baseline rank-one recognition rates of individual subsets with different variations in 3D FR on Lock3DFace. Specifically, the proposed method achieves the accuracy of 74.12% on Probe_Set_1, which includes the expressive face models with the frontal pose. This performance is similar to the ones around 78% achieved by ICP [6] [11] on the FRGC v2.0 dataset, where the expression change is the major challenge as well. With respect to the face models with various head poses in Probe_Set_2, the performance is largely degraded to 18.63%, and this sharp decrease is mainly due to serious data missing in many face samples caused by severe self-occlusion. Regarding Probe_Set_3 composed by the face models that are occluded by external objects, the accuracy is 28.57%, and

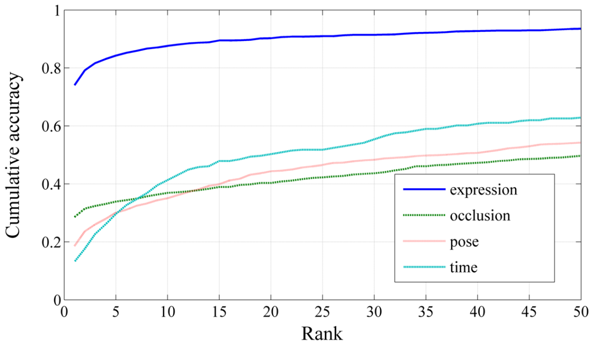

Figure 6. Cumulative Match Characteristic (CMC) curves of different scenarios on the Lock3DFace database.

this limited score is the result of the incompetency of holistic ICP matching on partially impaired data. In the case of Probe_Set_4 concerning the problem of time lapse, the precision on the face samples captured in Session-2, 7 months after Session-1, is 13.17%. It should be noted in the second session, face samples with other variations besides time lapse are taken into consideration, and the distance between the volunteers and the Kinect sensor varies as well. The 3D face models generated by KinectFusion are thus of different resolutions, leading to the even deteriorative result. The Cumulative Match Characteristic (CMC) curves of those scenarios are depicted in Fig.6.

Table 4. Baseline results of 3D FR of different scenarios on the Lock3DFace database.

Variation	Rank-1 RR	Gallery Set	Probe Set
Expression	74.12%	Neutral-frontal (509)	Probe_Set_1 (1287)
Pose	18.63%		Probe_Set_2 (1014)
Occlusion	28.57%		Probe_Set_3 (1004)
Time	13.17%		Probe_Set_4 (1352)
Average	34.53%		Probe_Set (4657)

In general, the average recognition rate on Lock3DFace is 34.53%, which needs significant improvement in the future. From another point of view, we can see that it is still challenging in low cost 3D FR, and Lock3DFace currently provides the best benchmark for this issue.

6. Conclusion and Future Work

We summarize the current low cost 3D face database and deliver a new one of large-scale captured by Kinect, namely Lock3DFace, which makes a significant improvement both in quantity and challenge for FR. Furthermore, we present the standard protocol in low cost 3D FR on this database, and report the baseline results using the method, which combines KinectFusion based model reconstruction and ICP based matching.

Regarding the future work, we will further evaluate more existing method on the Lock3DFace database and seek possible solutions to deal with the tough challenges especially on the subsets of pose, occlusion, and time lapse. Moreover, multi-modal 2D+3D FR approaches will also be investigated to further improve the performance on this dataset.

Acknowledgement

This work was supported in part by the Hong Kong, Macao, and Taiwan Science and Technology Cooperation Program of China (Grant No. L2015TGA9004), in part by the National Natural Science Foundation of China (Grant No. 61540048), and in part by the Fundamental Research Funds for the Central Universities.

References

[1] http://biometrics.idealtest.org/. 2

[2] S. Berretti, A. Del Bimbo, and P. Pala. Superfaces: A super-resolution model for 3d faces. In *European Conference on Computer Vision Workshops and Demonstrations*, pages 73–82. Springer, 2012. 2

[3] S. Berretti, P. Pala, and A. Del Bimbo. Face recognition by super-resolved 3d models from consumer depth cameras. *IEEE Transactions on Information Forensics and Security*, 9(9): 1436–1449, 2014. 2

[4] P. J. Besl and H. D. Mckay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2): 239–256, 1992. 6

[5] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: a 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3): 413–425, 2014. 2, 3

[6] K. I. Chang, K. W. Bowyer, and P. J. Flynn. Adaptive rigid multi-region selection for handling expression variation in 3d face recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 157–157. IEEE, 2005. 6

[7] K. I. Chang, K. W. Bowyer, and P. J. Flynn. Multiple nose region matching for 3d face recognition under varying facial expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10): 1695–1700, 2006. 1

[8] A. Colombo, C. Cusano, and R. Schettini. Umb-db: A database of partially occluded 3d faces. In *International Conference on Computer Vision Workshops*, pages 2113–2119. IEEE, 2011. 2

[9] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Annual Conference on Computer Graphics and Interactive Techniques*, pages 303–312. ACM, 1996. 5

[10] N. Erdogmus and S. Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6. IEEE, 2013. 2, 3

[11] B. Gökberk, H. Dutağaci, A. Ulaş, L. Akarun, and B. Sankur. Representation plurality and fusion for 3-d face recognition.

IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 38(1): 155–173, 2008. 6

[12] M. Hernandez, J. Choi, and G. Medioni. Laser scan quality 3-d face modeling using a low-cost depth camera. In *European Signal Processing Conference*, pages 1995–1999. IEEE, 2012. 5

[13] D. Huang, M. Ardabilian, Y. Wang, and L. Chen. 3-d face recognition using elbp-based facial description and local feature hybrid matching. *IEEE Transactions on Information Forensics and Security*, 7(5): 1551–1565, 2012. 1

[14] D. Huang, W. Ben Soltana, M. Ardabilian, Y. Wang, and L. Chen. Textured 3d face recognition using biological vision-based facial representation and optimized weighted sum fusion. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2011. 1

[15] D. Huang, K. Ouji, M. Ardabilian, Y. Wang, and L. Chen. 3d face recognition based on local shape patterns and sparse representation classifier. In *International Conference on Multimedia Modeling*, pages 206–216. Springer, 2011. 6

[16] D. Huang, J. Sun, X. Yang, D. Weng, and Y. Wang. 3d face analysis: Advances and perspectives. In *Chinese Conference on Biometric Recognition*, pages 1–21. Springer, 2014. 1

[17] T. Huynh, R. Min, and J.-L. Dugelay. An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In *Asian Conference on Computer Vision Workshops*, pages 133–145. Springer, 2013. 2

[18] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Annual on UserInterface Software and Technology*, pages 559–568. ACM, 2011. 5

[19] B. Y. Li, A. Mian, W. Liu, and A. Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Workshop on Applications of Computer Vision*, pages 186–192. IEEE, 2013. 2, 3, 5

[20] H. Li, D. Huang, J.-M. Morvan, Y. Wang, and L. Chen. Towards 3d face recognition in the real: A registration-free approach using fine-grained matching of 3d keypoint descriptors. *International Journal of Computer Vision*, 113(2): 128–142, 2015. 1

[21] P. Liu, Y. Wang, D. Huang, Z. Zhang, and L. Chen. Learning the spherical harmonic features for 3-d face recognition. *IEEE Transactions on Image Processing*, 22(3): 914–925, 2013. 1

[22] S. Luuk. Fast and accurate 3d face recognition using registration to an intrinsic coordinate system and fusion of multiple region classifiers. *Inteurnal of computer Vision*, 93(3): 389–414. 1

[23] A. S. Mian, M. Bennamoun, and R. Owens. An efficient multimodal 2d-3d hybrid approach to automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11): 1927–1943, 2007. 1

[24] R. Min, J. Choi, G. Medioni, and J.-L. Dugelay. Real-time 3d face identification from a depth camera. In *International Conference on Pattern Recognition*, pages 1739–1742. IEEE, 2012. 1

[25] R. Min, N. Kose, and J.-L. Dugelay. Kinectfacedb: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(11): 1534–1548, 2014. 2

[26] N. Mori, T. Suzuki, and S. Kakuno. Noise of acoustic doppler velocimeter data in bubbly flows. *Journal of Engineering Mechanics*, 133(1): 122–125, 2007. 5

[27] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011. 5

[28] 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. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 947–954. IEEE, 2005. 1, 2

[29] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *COST 2101 Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008. 1, 2

[30] D. Smeets, P. Claes, J. Hermans, D. Vandermeulen, and P. Suetens. A comparative study of 3-d face recognition under expression variations. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(5): 710–727, 2012. 1

[31] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014. 1

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1701–1708. IEEE, 2014. 1

[33] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *International Conference on Computer Vision*, pages 839–846. IEEE, 1998. 5

[34] V. Vijayan, K. W. Bowyer, P. J. Flynn, D. Huang, L. Chen, M. Hansen, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris. Twins 3d face recognition challenge. In *International Joint Conference on Biometrics*, pages 1–7. IEEE, 2011. 2

[35] Y. Wang, J. Liu, and X. Tang. Robust 3d face recognition by local shape difference boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10): 1858–1870, 2010. 1

[36] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *International Conference on Automatic Face and Gesture Recognition*, pages 1–6. IEEE, 2008. 2

[37] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition*, pages 211–216. IEEE, 2006. 1, 2