Revisiting Depth-based Face Recognition from a Quality Perspective

Zhenguo Hu, Qijun Zhao
College of Computer Science,
Sichuan University, Chengdu, China
qjzhao@scu.edu.cn

Feng Liu
College of Computer Science
and Software Engineering,
Shenzhen University, Shenzhen, China
feng.liu@szu.edu.cn

Abstract

Face recognition using depth data has attracted increasing attention from both academia and industry in the past five years. Despite the large number of depth-based face recognition methods in the literature, high quality data are usually required for high recognition accuracy. In this paper, we measure the quality of 3D face data in terms of resolution and precision, and evaluate how the accuracy of three deep face recognition models varies on several benchmark databases as the facial depth data resolution changes from dense to sparse and as the precision changes from high to low. From the experimental results, several observations are made. (i) Given a high precision, a low resolution of $3K$ *is sufficient to represent a 3D face; when the precision decreases, using higher resolutions can benefit face recognition, but the recognition accuracy becomes saturated as the resolution reaches* $10K$*. (ii) Depth precision is more critical than resolution in depth-based face recognition, and a precision of* $1mm$ *is generally preferred as a good balance between accuracy and cost. (iii) The deep models trained with low-quality data perform more stable across data of different quality levels. We believe that these observations are beneficial for both depth sensor manufacturers and depth-based face recognition system developers.*

1. Introduction

Three-dimensional (3D) face recognition has been studied for several decades with a large variety of methods proposed [18, 13, 15, 16]. It is believed that 3D face data have intrinsic advantages over 2D face images in detecting presentation attacks and in providing additional discriminative features for face recognition [3]. Yet, 3D face recognition had not gained popularity in real-world applications until Apple Inc. released its iPhone X with TrueDepth camera and Face ID in 2017. One reason is due to that the scanners used for acquiring 3D faces in previous studies are mostly bulky and expensive, and are thus not feasible in practical scenarios, though previous studies [18, 13, 15] obtained very high recognition accuracy by using the captured high quality 3D face data (see Table 1).

The emergence of low-cost RGB-D sensors, such as Kinect [26] and RealSense [9], makes it possible to capture 3D faces more efficiently and more cost-effectively. Many attempts [1, 8, 14, 10, 27] have been made in the past years to develop practical face recognition systems based on RGB-D sensors. As shown in Table 1, with depth images as auxiliary information, researchers [27, 14] show that face recognition accuracy can be improved compared with using only RGB images. However, the accuracy achieved by using depth images captured by low-cost RGB-D sensors [1, 8, 14, 27] is still much lower than that by using 3D faces captured by 3D scanners [18, 15]. This is because the quality of the depth images captured by low-cost RGB-D sensors is generally poor (see Fig. 1).

Figure 1. Depth images captured by different devices or under different conditions show different quality levels. From left to right: Facial depth images captured by Konica Minolta Vivid 910 [17] and 3dMD [23] in lab, Kinect II in lab [28], RealSense in lab and in the wild.

Inspired by the success of deep learning in computer vision tasks including face recognition with RGB images, some researchers proposed deep networks either for preprocessing facial depth images [7] or for learning effective depth feature representations of faces [7, 27]. They reported impressive improvement on face recognition accuracy, and demonstrated the potential of facial depth images as promising identity evidence. Yet, from the viewpoint of practical applications, there are still many open issues: e.g., How many points should be used to represent a 3D face? How

precise should the depth values be for reliable face recognition? Is it more significant to increase the number of points in the point cloud of a 3D face or to improve the precision of its depth values? How will a depth-based face recognition model generalize across data of different quality levels? To answer these questions, it is highly demanded to further investigate the capacity of depth-based face recognition technology under various conditions such that the technology can be deployed in a more effective way. This paper thus provides a revisit to depth-based face recognition from a quality perspective with the aim of assessing the impact of facial depth image quality on face recognition accuracy.

To this end we focus on two intrinsic factors affecting the quality of 3D face data, i.e., resolution and precision. Resolution (also known as density) refers to the number of points used to represent a 3D face, and precision refers to the measuring accuracy of depth values (in terms of millimeter or mm). Table 1 summarizes these two quality metrics of 3D face data in different databases. In this paper, we conduct comprehensive face recognition experiments by using three well-known deep learning based models on five databases that are constructed with different 3D/RGB-D sensors. Particularly, we quantitatively evaluate the potential of low-quality facial depth images in identity recognition with respect to both resolution and precision of the underlying 3D face data.

According to the experimental results, several observations are made. i) Given a high precision, a low resolution of $3K$ is sufficient to represent a 3D face; when the precision decreases, using higher resolutions can benefit face recognition, but the recognition accuracy becomes saturated as the resolution reaches $10K$. ii) Precision is more critical than resolution in depth-based face recognition, and a precision of $1mm$ is generally preferred as an acceptable tradeoff between accuracy and cost. iii) Face recognition models trained on low quality depth images usually generalize better than models trained on high quality depth images. We believe that these observations can provide helpful references for manufacturers of depth sensors and developers of depth-based face recognition systems.

The rest of this paper is organized as follows. Section II introduces the databases and face recognition models used in this study. Section III introduces in detail the evaluation protocols. Section IV presents the obtained experimental results along with analysis and discussion. Section V finally concludes the paper with suggested future research directions.

2. Databases and Face Recognition Models

2.1. Databases

We use five databases in this paper, two of which are constructed with high-cost 3D scanners and the other three with low-cost RGB-D sensors. Below, we introduce the detail of these databases.

FRGC v2 [17] consists of $4,007$ 3D facial scans of 466 subjects acquired by using a laser 3D scanner, i.e., Konica Minolta Vivid 910. These 3D scans have relatively high resolution and precision. Specifically, their resolution ranges from $50K$ to $170K$, and their precision is about $0.1mm$. They are captured at frontal pose and with limited expression variations of low intensity. FRGC v2 is one of the most widely used benchmark databases in 3D face recognition research. In this paper, we will use it to generate depth images of varying resolutions and precisions for both training and testing.

BU3DFE [23] contains 3D faces of 100 subjects with different expressions, including neutral expression and six types of universal expressions (i.e., happiness, anger, sadness, surprise, fear, and disgust) at four intensity levels. The 3D faces are acquired by using a high-cost 3D scanner at a resolution around $8K$ and a precision about $0.2mm$. BU3DFE is among the most widely used benchmark databases for 3D facial expression recognition. In this paper, we generate depth images from the 3D faces of neutral and first-level happy and sad expressions in BU3DFE as test data to evaluate the generalization ability of face recognition models.

BUAA database [28], also known as Lock3DFace, captures face data by using the low-cost RGB-D sensor Kinect II in lab. It contains totally $5,711$ RGB-D video sequences of 509 Chinese subjects with variations in pose, expression, and occlusion. During acquisition, the subjects are asked to rotate their heads in both pitch and yaw directions by up to 90°, and display the six types of universal expressions at low intensity. The resolution and precision of the obtained 3D face data are $20K$ and $\geq 2mm$, and hence Lock3DFace is a low quality RGB-D face database. This database is established particularly for the purpose of evaluating the performance of face recognition with low-cost RGB-D sensors.

SCU-RGBD is a low quality RGB-D face database collected in lab by ourselves with the low-cost RGB-D sensor RealSense [9]. It contains 900 RGB-D video sequences of 247 Chinese subjects, who are asked to rotate their heads in yaw direction by -90° to $+90^\circ$ with neutral, surprise and smile expressions under varying illuminations. 3D faces in this database have a resolution of $45K$ and a precision of $\geq 2mm$. We will release this database in the public domain for research usage.

RGBD-W is also collected by ourselves with the RealSense sensor. Unlike SCU-RGBD, this database is constructed in the wild rather than in lab. Specifically, we mounted the RealSense sensor on a gate at the entrance of a railway station. When a person passed through the gate, he/she had to stand in front of the gate and had his/her identity card and ticket checked. RGB-D face data was col-

Table 1. Benchmark databases and state-of-the-art recognition accuracy on them when using depth images only, RGB images only, or both depth and RGB images. SCU-RGBD and RGBD-W are our collected databases.

Databases	No. of Subjects	Devices	Resolution	Precision (mm)	Rank-1 Identification Rate Using		
					Depth	RGB	Depth+RGB
FRGC v2 [17]	466	Vivid 910	$60K$	0.1	99.6% [18]	–	–
BU-3DFE [23]	100	3dMD	$8K$	0.2	99.3% [15]	–	–
BUAA [28]	509	Kinect II	$20K$	≥ 2	66.0% [27]	92.5% [27]	93.2% [27]
SCU-RGBD	200	RealSense	$45K$	≥ 2	73.0%	–	–
RGBD-W	$2,239$	RealSense	$45K$	≥ 2	64.0%	94.7%	–
IIIT-D [5]	106	Kinect I	$13K$	$2-4$ [26]	26.8% [1]	99.0% [27]	98.7% [1]
CurtinFaces [8]	52	Kinect I	$13K$	$2-4$	72.5% [8]	87.0% [8]	91.3% [8]
Eurecom [14]	52	Kinect I	$13K$	$2-4$	69.7% [14]	94.6% [14]	96.3% [27]

lected during that time. We finally acquired 100 RGB-D face images for each of $2,293$ subjects. Neither illumination nor expression/pose of the subjects was controlled in this database. RGBD-W is used to evaluate the performance of depth-based face recognition models for in-the-wild applications.

Example depth images from the above five databases are shown in Fig. 1. Obviously, the quality of the depth images in FRGC v2 and BU-3DFE is much better than that of the depth images in the other three databases in terms of precision. As for low-cost depth sensors, it seems that RealSense can capture more detailed depth information than Kinect II, and acquisition under uncontrolled conditions usually results in even lower quality depth images.

2.2. Face Recognition Models

Three deep face recognition models, Lightened CNN [24], CASIA-Net [25] and SphereFace [11], are considered in this study. All are relatively light-weight models. This enables us not only to fine tune the models but also to train them from scratch by using relatively small data sets of facial depth images that were available to us. Note that although some models in our experiments obtain state-of-the-art results, our main goal in this paper is to assess the impact of quality factors rather than to promote the state-of-the-art of depth-based face recognition. Therefore, we do not employ complex or very deep models like VGG [20] and GoogleNet [21].

The structure of Lightened CNN is the same as in [24]. Its input image size is 128×128, and the $256-$dimensional output of $FC1$ layer is taken as the extracted feature. For CASIA-Net, motivated by [22], we add batch normalization [12] and exponential linear unit [2] after each convolutional layer. The input image size is changed from 100×100 to 96×96, and the $320-$dimensional output of $Pool5$ layer is taken as the extracted feature. For SphereFace, we employ SphereFace-20 as defined in [11]. Its input image size is 112×96, and the $512-$dimensional output of $FC1$ is taken as the extracted feature. For all the three deep models, $cosine$ similarity is employed to measure the similarity between the extracted features of different facial depth images.

3. Evaluation Protocols

In order to evaluate the impact of depth data quality (in terms of resolution and precision in this paper) on face recognition accuracy, we conduct evaluation experiments using both synthetic and real-world data. The synthetic data, including 3D faces of varying resolution and precision, are generated from the high quality data in FRGC v2. When training the deep models, we also augment the 3D face data via rotating them by different amounts to generate multi-pose faces. The 3D face data are mapped to 2D planes via weak perspective projection, resulting in depth images that are required by the deep face recognition models. The depth images, after necessary pre-processing, are organized into training and testing subsets to assess the performance of the deep face recognition models. Below we introduce in detail our evaluation protocols.

3.1. Generating Depth Images of Varying Quality Levels

Using a set of facial depth images with systematic variations in resolution and precision is very helpful for assessing the impact of these quality factors. However, as introduced in the last section, existing databases can not provide such data. Therefore, we synthesize facial depth images of varying quality from the high quality 3D faces in FRGC v2. For this purpose, we choose for each of the 466 subjects the frontal 3D face scan with neutral expression as the original 3D face. All the chosen original 3D faces have their nose tips aligned in a common coordinate system, and are then cropped by using a sphere whose center is at the nose tip and radius is set as $120mm$, resulting in 3D faces whose resolution is in between $35K$ and $85K$. These 3D faces are further down-sampled by re-sampling the point clouds

at different resolutions, including $20K$, $15K$, $10K$, $5K$ and $3K$ in our experiments. To further augment the pose variations in the data sets, we rotate the 3D faces along yaw direction by $\pm 10°$, $\pm 20°$, $\pm 30°$, $\pm 40°$ and $\pm 45°$, and along pitch and roll direction by $\pm 10°$ and $\pm 15°$.

To simulate 3D faces of different precisions, we add random Gaussian noises to the depth values (i.e., $z-$coordinates) of the 3D faces. In our experiments, we consider two types of Gaussian noises whose means are zero and standard deviations are 10 and 20. Given that the precision of the original data is $0.1mm$, the obtained 3D faces with these two Gaussian noises have approximate depth precisions of $1mm$ and $2mm$, respectively. These 3D faces are finally projected to 2D planes via weak perspective projection, resulting in 275 depth images per quality level (i.e., at certain resolution and precision) per subject. Figure 2 shows some obtained 3D face shapes and corresponding depth images at different quality levels.

3.2. Data Preprocessing and Organization

The depth images are first aligned based on the five landmarks of left and right eye centers, left and right mouth corners, and nose tip on them such that the two eyes locate on a horizontal line, and then cropped and resized to the specific input size required by different models (refer to Sec. II.B). Here, the landmarks on the depth images in FRGC v2 and BU3DFE are directly obtained from the landmarks annotated on the source 3D faces, while the landmarks for the other databases are either automatically detected by using MTCNN [29] or manually marked (if MTCNN fails). Note that since depth images are registered with their corresponding RGB/NIR images, the landmarks are detected by applying the publicly available MTCNN model to the RGB/NIR images. Finally, the pixel values on each depth image are normalized to the interval $[0, 1]$ via min-max normalization.

As introduced above, FRGC v2, BUAA, SCU-RGBD and RGBD-W are used for both training and testing. Hence, we divide each of them into training, validation and testing subsets. For FRGC v2, the first 314 subjects are chosen for training; after shuffling their $86,350$ facial depth images, the first $15,000$ images are picked as validation subset, and the remaining images are used as training subset. The images of the other 152 subjects in FRGC v2 compose the testing subset, among which a frontal depth image per subject is chosen as gallery and the rest images are used as probes.

In BUAA, we randomly select 330 subjects for training and validation, among whose images $80K$ images are randomly chosen as training subset and the remaining $20K$ images as validation subset. For the other 179 subjects in the testing subset, one of the frontal neutral images is chosen per subject to form the gallery, while the other frontal neutral images compose one probe subset BUAA-NU, and all the rest images with varying poses and expressions compose another probe subset BUAA-ALL.

In SCU-RGBD, 147 subjects are randomly chosen for training and validation, and the other 100 subjects for testing. More specifically, the training and validation subsets contain $45K$ and $5K$ images, respectively. In the testing subset, the gallery consists of one frontal neutral image per subject, and the probe consists of all the remaining images.

In RGBD-W, $1,293$ subjects are randomly chosen for training and validation, and the other $1,000$ subjects for testing. As a result, the training and validation subsets contain $200K$ and $20K$ images, respectively. As in RGBD-W, the gallery in the testing subset consists of one frontal neutral image per subject, while the probe consists of all the remaining images.

The BU3DFE database is used only for testing. One frontal neutral image of each subject is chosen as gallery, while all the other images are used as probe. Note that the 3D faces in BU3DFE are augmented via rotation in the same way as in FRGC v2 (refer to Sec. 3.1).

3.3. Model Training and Performance Metrics

We train the three deep models in two different ways, training from scratch and fine-tuning a pre-trained model. When training Lightened CNN and SphereFace from scratch that are implemented on Caffe, we set dropout rate to 0.7, momentum and weight decay to 0.9 and $5e-4$, and initial learning rate to $1e-3$, which is gradually reduced to $1e-5$. As for CASIA-Net, we implement it on TensorFlow [4]. When training it from scratch, the model is initialized by a zero-centered normal distribution with a standard deviation of 0.02, and optimized by using the Adam optimizer [6]. The learning rate is first set as $2e-4$ and updated to $1e-4$ when the training is saturated. We finish the training when the loss on validation subset is below $1e-3$. When fine-tuning Lightened CNN and SphereFace, the pre-trained models provided by [24] and [11] are used, and the learning rate is set to $1e-5$. When fine-tuning CAISA-Net, we first train the model by using the RGB face images in CASIA-Webface [25], and then fine-tune the obtained model on facial depth images with a learning rate of $1e-5$. We evaluate the depth-based face recognition performance of these deep models in identification mode, and compare their rank-1 identification rates when depth images of different quality levels are used.

4. Results and Discussion

In this section, we first report the evaluation results on synthetic data generated from FRGC v2, and then report the results on real-world data. Discussion will be presented along with the results.

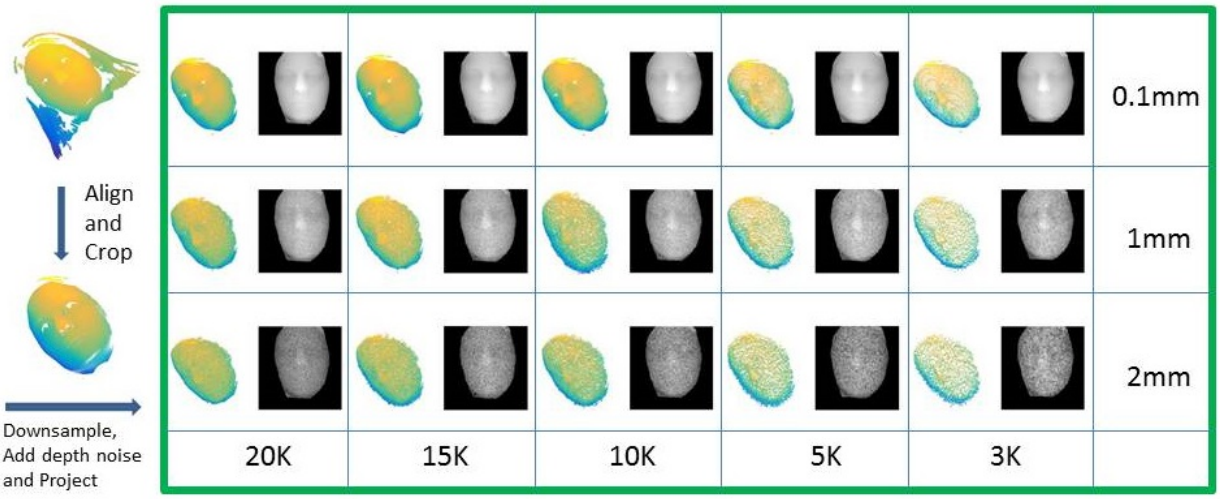

Figure 2. Procedure of generating depth images of various quality levels. Given a high quality 3D face, it is first aligned and cropped, then downsampled to different resolutions (including $20K$, $15K$, $10K$, $5K$ and $3K$) and added with depth noise (resulting in a precision of $0.1mm$, $1mm$ or $2mm$), and finally projected to 2D plane to form depth images. Note that the pose augmentation step is not shown for the sake of highlighting the simulation of different quality levels.

4.1. Results on Synthetic Data

Experiments using synthetic data consider both homogeneous and heterogeneous face recognition. In homogeneous face recognition, training and testing data share the same resolution and precision, i.e., the same quality level. In heterogeneous face recognition, training and testing data could have different resolutions or precisions. Below, we report the respective results.

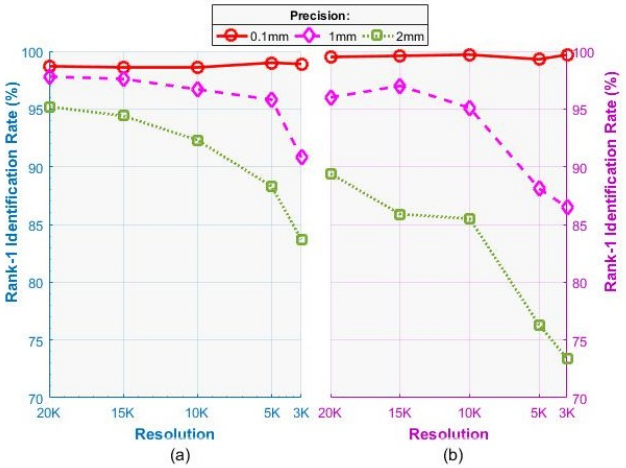

Figure 3. Rank-1 identification rates of (a) Lightened CNN and (b) CASIA-Net in homogeneous face recognition on the depth images generated from FRGC v2 at different quality levels.

4.1.1 Homogeneous Face Recognition

Figure 3 summarizes the results of Lightened CNN and CASIA-Net for homogeneous face recognition. In this experiment, all the models are trained from scratch. As can be seen, given a specific precision, the recognition accuracy is in general decreases as the resolution drops from $20K$ to $3K$. Given a specific resolution, similar trends can be observed for the impact of precision on the recognition accuracy. However, when one quality factor is at high level (e.g., a resolution of $20K$ or a precision of $0.1mm$), the accuracy degradation due to the decline of the other quality factor becomes less serious. Especially, for example, at the precision of $0.1mm$, the accuracy of both models remains almost stable when the resolution changes; in contrast, at the precision of $2mm$, their accuracy both decreases obviously (about 12% for Lightened CNN and 18% for CASIA-Net) when the resolution is reduced from $20K$ to $3K$.

On the one hand, these results reveal the potential benefit of enhancing depth image quality. A real-world practice of this idea is Kinect fusion [19], which aims to generate 3D data with higher resolution (and possibly higher precision also) by fusing multiple continuous frames captured by Kinect. On the other hand, the gain of quality improvement becomes marginal at certain levels. Taking the CASIA-Net as an example, at the precision of $1mm$, the contribution of improved resolution seems saturated at the resolution of $10K$.

To sum up, when the precision is high (i.e., $0.1mm$), a relatively low resolution (i.e., $3K$) is sufficient to represent a 3D face. As the precision decreases, higher resolutions

will be needed to maintain a relatively high recognition accuracy. However, when the resolution is as high as $10K$, the gain of increased resolutions becomes marginal.

4.1.2 Heterogeneous Face Recognition

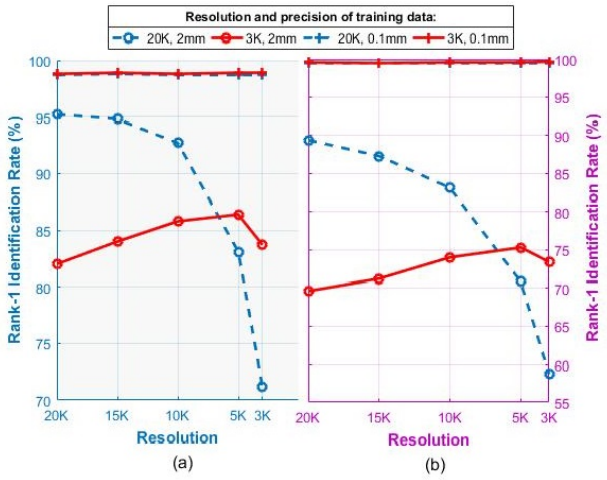

Figure 4. Rank-1 identification rates of (a) Lightened CNN and (b) CASIA-Net in heterogeneous face recognition on the depth images generated from FRGC v2. The training and testing data have the same precision, but may differ in resolution.

Figure 4 and Table 2 present the recognition accuracy of Lightened CNN and CASIA-Net for heterogeneous face recognition. In Figure 4, we assume that the training and testing data have the same precision (i.e., either $0.1mm$ or $2mm$), and may differ in resolution. It can be clearly seen from the results that when the precision is as high as $0.1mm$, variations in resolution do not have apparent effect on the recognition accuracy. This again demonstrates that given a sufficiently high precision, a resolution of $3K$ would be sufficient for representing a 3D face. However, when the data have a low precision of $2mm$, the accuracy changes much more obviously on probes of different resolutions.

In Table 2, we assume that the training and testing data have the same resolution (i.e., either $20K$ or $3K$), and may differ in precision. Unlike the results in Figure 4, obvious variations in recognition accuracy can be observed here when training and testing depth images have different precisions. Moreover, the effect of precision heterogeneity on face recognition accuracy is more striking on data of lower resolution (i.e., $3K$). These results reveal the relatively more significant importance of precision than resolution for reliable depth-based face recognition.

According to Figure 4 and Table 2, it is also worth mentioning that the models trained with relatively lower quality depth data (e.g., resolution of $3K$ and precision of $1mm$) perform more stable when applied to probes of different

Table 2. Rank-1 identification rates (%) of Lightened CNN / CASIA-Net in heterogeneous face recognition on the depth images generated from FRGC v2. The training and testing data have the same resolution, but may differ in precision.

Precision of Training Data	Precision of Testing Data		
	0.1mm	1mm	2mm
	Resolution of both training and testing data: 20K		
0.1mm	98.7 / 99.5	93.5 / 84.4	70.2 / 32.1
1mm	83.1 / 81.5	97.8 / 96.0	83.9 / 79.5
2mm	39.0 / 58.5	66.2 / 80.0	95.2 / 89.4
	Resolution of both training and testing data: 3K		
0.1mm	98.8 / 99.7	58.4 / 28.2	25.6 / 8.5
1mm	78.6 / 60.0	90.8 / 86.5	66.7 / 46.6
2mm	33.6 / 35.7	53.5 / 55.0	83.7 / 73.4

Table 3. Rank-1 identification rates (%) of Lightened CNN / CASIA-Net in heterogeneous face recognition on real-world data sets.

Training Data Sets	Testing Data Sets			
	BU3DFE	BUAA	SCU-RGBD	RGBD-W
FRGC v2	37.5 / 51.9	–	–	–
BUAA	–	–	46.3 / 55.6	12.0 / 30.0
RGBD-W	–	56.2 / 57.2	51.5 / 46.9	–

quality levels. This demonstrates their better generalization ability, although their accuracy on high quality data might be not as high as that of the models also trained with high quality data. To further verify this observation, we evaluate the model trained with synthetic data of resolution $10K$ and precision $0.1mm$ and the model trained with BUAA and RGBD-W on BU3DFE, SCU-RGBD and RGBD-W. The obtained results as shown in Table 3 lead to the same observation. The accuracy of both Lightened CNN and CASIA-Net models trained on FRGC v2, when tested on BU3DFE, decreases by about 50%. On the contrary, when they are trained on low quality depth data, their accuracy degradation on other data sets is within 25%. Note that the scale of RGBD-W is much larger than that of BUAA, which leads to a substantial decrease of recognition accuracy.

4.2. Results on Real-World Data

We also evaluate the accuracy of the considered deep face recognition models on real-world data. The results are shown in Table 4. As can be seen, the deep models achieve reasonably good accuracy on the real-world data collected with low-cost RGB-D sensors (i.e., Kinect or RealSense), compared with the state-of-the-art results in Table 1 and the results on synthetic data. Also, the decrease of performance on synthetic data from high quality to low quality is consistent with the one on real-world data, which demonstrates the reasonability of the data generation.

Comparing the results obtained by different training

Table 4. Rank-1 identification rates (%) of Lightened CNN, CASIA-Net and SphereFace in homogeneous face recognition on different data sets when the models are either trained from scratch or first pre-trained on RGB face images and then fine-tuned with depth images.

Deep Models	Databases	Identification Rate (%)	
		From Scratch	Fine-tuned
Lightened CNN	BUAA-NU	78.43	80.0
	BUAA-ALL	57.22	57.75
	SCU-RGBD	66.5	71.9
	RGBD-W	58.7	56.7
CASIA-Net	BUAA-NU	90.9	91.0
	BUAA-ALL	76.7	73.0
	SCU-RGBD	72.7	73.0
	RGBD-W	64.0	63.5
SphereFace	BUAA-NU	73.0	77.6
	BUAA-ALL	50.8	54.0
	SCU-RGBD	52.3	55.8
	RGBD-W	54.2	55.7

methods, we obverse that using RGB face images to pre-train the deep models and then using facial depth images to fine-tune the pre-trained models can effectively improve the face recognition accuracy in most cases. Comparing the results across different databases, we find that variations in pose and expression are other quality factors than resolution and precision that could significantly affect the depth-based face recognition accuracy. Comparing the results across different deep models, we can see that although SphereFace performs best for RGB-based face recognition, its performance on depth images is worse than that of the other two deep models. A possible reason is because its complexity is one order of magnitude higher than that of the other models and the number of subjects in the training data is quite small with respect to the number of its parameters.

5. CONCLUSIONS AND FUTURE WORK

This paper presents a revisit to depth-based face recognition from the quality perspective. Three deep face recognition models are evaluated on both synthetic and real-world data from five benchmark databases. The comprehensive evaluation results demonstrate that (i) a low resolution of $3K$ is sufficient to represent a 3D face at a high precision, and as the precision decreases, improving the resolution can significantly benefit face recognition until the resolution is as high as $10K$, (ii) compared with resolution, depth precision can more significantly affect the accuracy of depth-based face recognition systems, and (iii) training the face recognition models with low quality data is helpful in improving their generalization ability across data of different quality levels. One limitation of this study is due to the relatively small-scale datasets. Yet, we believe that the observations made in our evaluation are helpful for both researchers and practitioners in the field of depth-based / 3D face recognition.

Our evaluation results also suggest the following valuable research directions in depth-based face recognition.

- As precision is more critical than resolution, it deserves to explore software-based methods for enhancing the precision of the depth data captured by low-cost RGB-D sensors.
- Considering the quality variations commonly occurring in practical applications, it is of significant importance to develop effective depth-based face recognition methods that are robust across data of different quality levels, and especially for in-the-wild applications.
- Existing databases of facial depth data in the public domain are relatively small-scale. It is highly demanded to establish large-scale depth-based face databases to enable large-scale evaluation of relevant technology as well as large-scale verification of relevant conclusions.

6. ACKNOWLEDGEMENT

This work is supported by the National Key Research and Development Program of China (2017YFB0802300), the National Natural Science Foundation of China (61773270, 61703077) and the Shenzhen Fundamental Research fund (JCYJ20180305125822769).

References

[1] A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa. RGB-D face recognition via learning-based reconstruction. In *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*, pages 1–7, 2016.

[2] D. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, abs/1511.07289, 2015.

[3] N. Erdogmus and S. Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, Arlington, VA, USA, September 29 - October 2, 2013*, pages 1–6, 2013.

[4] F. Giannini, V. Laveglia, A. Rossi, D. Zanca, and A. Zugarini. Neural networks for beginners. A fast implementation in matlab, torch, tensorflow. *CoRR*, abs/1703.05298, 2017.

[5] G. Goswami, M. Vatsa, and R. Singh. RGB-D face recognition with texture and attribute features. *IEEE Trans. Information Forensics and Security*, 9(10):1629–1640, 2014.

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[7] Y. Lee, J. Chen, C. W. Tseng, and S. Lai. Accurate and robust face recognition from RGB-D images with a deep learning approach. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

[8] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna. Face recognition based on kinect. *Pattern Anal. Appl.*, 19(4):977–987, 2016.

[9] C. Lin, C. Wang, H. Chen, W. Chu, and M. Y. Chen. Realsense: directional interaction for proximate mobile sharing using built-in orientation sensors. In *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, pages 777–780, 2013.

[10] H. Liu, F. He, Q. Zhao, and X. Fei. Matching depth to RGB for boosting face verification. In *Biometric Recognition - 12th Chinese Conference, CCBR 2017, Shenzhen, China, October 28-29, 2017, Proceedings*, pages 127–134, 2017.

[11] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6738–6746, 2017.

[12] Y. Ma and D. Klabjan. Convergence analysis of batch normalization for deep neural nets. *CoRR*, abs/1705.08011, 2017.

[13] A. S. Mian, M. Bennamoun, and R. A. Owens. An efficient multimodal 2d-3d hybrid approach to automatic face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(11):1927–1943, 2007.

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

[15] O. Ocegueda, G. Passalis, T. Theoharis, S. K. Shah, and I. A. Kakadiaris. UR3D-C: linear dimensionality reduction for efficient 3d face recognition. In *2011 IEEE International Joint Conference on Biometrics, IJCB 2011, Washington, DC, USA, October 11-13, 2011*, pages 1–6, 2011.

[16] P. Perakis, T. Theoharis, G. Passalis, and I. A. Kakadiaris. Automatic 3d facial region retrieval from multi-pose facial datasets. In *Eurographics Workshop on 3D Object Retrieval, Munich, Germany, 2009. Proceedings*, pages 37–44, 2009.

[17] P. J. Phillips, P. J. Flynn, W. T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. J. Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 947–954, 2005.

[18] C. C. Queirolo, L. Silva, O. R. P. Bellon, and M. P. Segundo. 3d face recognition using simulated annealing and the surface interpenetration measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):206–219, 2010.

[19] M. W. Rahman and M. L. Gavrilova. Emerging EEG and kinect face fusion for biometrie identification. In *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017, Honolulu, HI, USA, November 27 - Dec. 1, 2017*, pages 1–8, 2017.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9, 2015.

[22] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1283–1292, 2017.

[23] Y. V. Venkatesh, A. A. Kassim, J. Yuan, and T. D. Nguyen. On the simultaneous recognition of identity and expression from BU-3DFE datasets. *Pattern Recognition Letters*, 33(13):1785–1793, 2012.

[24] X. Wu, R. He, Z. Sun, and T. Tan. A light CNN for deep face representation with noisy labels. *IEEE Trans. Information Forensics and Security*, 13(11):2884–2896, 2018.

[25] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.

[26] S. Zennaro, M. Munaro, S. Milani, P. Zanuttigh, A. Bernardi, S. Ghidoni, and E. Menegatti. Performance evaluation of the 1st and 2nd generation kinect for multimedia applications. In *2015 IEEE International Conference on Multimedia and Expo, ICME 2015, Turin, Italy, June 29 - July 3, 2015*, pages 1–6, 2015.

[27] H. Zhang, H. Han, J. Cui, S. Shan, and X. Chen. RGB-D face recognition via deep complementary and common feature learning. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 8–15, 2018.

[28] J. Zhang, D. Huang, Y. Wang, and J. Sun. Lock3dface: A large-scale database of low-cost kinect 3d faces. In *International Conference on Biometrics, ICB 2016, Halmstad, Sweden, June 13-16, 2016*, pages 1–8, 2016.

[29] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016.