KinectFaceDB: A Kinect Database for Face Recognition

Rui Min, Neslihan Kose, and Jean-Luc Dugelay, *Fellow, IEEE*

***Abstract*—The recent success of emerging RGB-D cameras such as the Kinect sensor depicts a broad prospect of 3-D data-based computer applications. However, due to the lack of a standard testing database, it is difficult to evaluate how the face recognition technology can benefit from this up-to-date imaging sensor. In order to establish the connection between the Kinect and face recognition research, in this paper, we present the first publicly available face database (i.e., KinectFaceDB[1]) based on the Kinect sensor. The database consists of different data modalities (well-aligned and processed 2-D, 2.5-D, 3-D, and video-based face data) and multiple facial variations. We conducted benchmark evaluations on the proposed database using standard face recognition methods, and demonstrated the gain in performance when integrating the depth data with the RGB data via score-level fusion. We also compared the 3-D images of Kinect (from the KinectFaceDB) with the traditional high-quality 3-D scans (from the FRGC database) in the context of face biometrics, which reveals the imperative needs of the proposed database for face recognition research.**

***Index Terms*—Database, Face recognition, Kinect.**

I. Introduction

RECENTLY, the emerging RGB-D cameras such as the Kinect sensor [1] have been successfully applied to many 3-D based applications. Thanks to its efficiency, low-cost, ease of RGB-D mapping, and multimodal sensing, the Kinect sensor has received vast amount of attention from diverse research communities [2], including but not limited in computer vision [3], computer graphics [4], [5], augmented reality (AR) [4], human–computer-interaction (HCI) [6], instrument measurement [7], and robotics [8]. For biometrics, directly inheriting from its application in body parts segmentation and tracking [9], a number of algorithms (based on the Kinect) have been proposed for gait recognition [10]–[12] and body anthropometric analysis [13]–[15]. However, the adoption of this powerful new sensor for face recognition has been mostly overlooked due to the lack of a standard database for testing, which greatly limits the development of new algorithms and applications in this thriving research field. Therefore, it is of great importance to provide a standard database for researchers to develop and test 3-D/multimodal (i.e., 2-D + 3-D) face recognition methods using the Kinect data, so as to establish the connection between the Kinect and face recognition research.

Face recognition [16], the least intrusive biometric technique from the acquisition point of view, has been applied to a wide range of commercial and law enforcement applications. In comparison to other popular biometric traits (such as fingerprint [17] and iris [18]), biometric recognition (identification/verification) based on face requires the least user cooperation and thus can be applied in many advanced conditions (e.g., in video surveillance). Although benchmark methods (e.g., Eigenface [19], Fisherface [20], and local binary patterns (LBP) [21]) report highly accurate results on well controlled datasets, the introduction of new face databases [for example, the face recognition technology (FERET) database [22], the face recognition vendor test (FRVT) database [23], and the face recognition grand challenge (FRGC) database [24] proposed by the National Institute of Standards and Technology (NIST)] challenges the standard techniques by deploying different variations (expressions, illuminations, poses, etc.), large number of subjects, as well as new data modalities (2-D, 3-D, video, etc.). As a consequence, the proposed databases significantly promoted the development of robust and reliable face recognition methods.

Recent surveys [25], [26] have suggested that face recognition methods exploiting 3-D cues are more efficient than 2-D-based methods in different aspects. For instance, 3-D shape information is illumination invariant; it can provide complementary information in addition to 2-D in classification; and faces with different poses can be aligned in 3-D via rigid registrations [27], [28]. However, most of the literature works conduct experiments using high-quality 3-D scans (e.g., 3-D faces in FRGC, which are captured by a digital laser scanner [29], with depth resolution of 0.1 mm within typical sensing range), which can lead to an unbalanced matching between 2-D and 3-D data in terms of acquisition efficiency and data accuracy. With respect to the acquisition efficiency, the capturing of a high-resolution RGB image normally takes less than 0.05 s, whereas the laser scanning of a face takes 9 s in average [24]. Therefore, high-quality 3-D face scanning needs careful user cooperation, which can significantly slow down noncooperative 2-D face recognition

Manuscript received March 31, 2013; revised October 29, 2013 and February 20, 2014; accepted April 24, 2014. This work was supported by the European Project Tabula Rasa (EU FP7). This paper was recommended by Associate Editor V. Piuri.

R. Min is with the Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA (e-mail: rmin@med.unc.edu).

N. Kose and J.-L. Dugelay are with the Department of Multimedia Communications, EURECOM, Les Templiers, Biot 06410, France (e-mail: kose@eurecom.fr; jld@eurecom.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2014.2331215

[1]Online at http://rgb-d.eurecom.fr

when integrating 3-D. In regards to the data accuracy, the measurement of an object with 10 cm depth along the z-axis needs 10 bits representation, whereas all intensity information are represented by 8 bits. Due to such imbalances, it is difficult to efficiently deploy 2-D and 3-D integrated face recognition system in practical scenarios.

Fortunately, the Kinect sensor overcomes above problems by providing both 2-D and 3-D data simultaneously at interactive rates. As a result, practical 3-D or 2-D + 3-D face recognition systems can be implemented for real-time and online processing [30]. However, in comparison to the high-quality laser scans, the quality of 3-D data captured by the Kinect is relatively low. In particular, it suffers from the problems of missing data in "blind points" [2], relatively low depth resolution, noise at large depth transitions (i.e., at boundaries), and spatial calibration/mapping of RGB and depth images [31], [32]. In face recognition, it is important to evaluate the 3-D data quality of the Kinect (using biometric metric) in comparison to high-quality laser scans. Moreover, efficient combination of 2-D and 3-D data captured from the Kinect is also an important issue to improve recognition performance.

In this paper, we present a standard database (i.e., the KinectFaceDB) to evaluate face recognition algorithms using the Kinect sensor. The proposed database consists of 936 shots of well-aligned 2-D, 2.5-D, and 3-D face data, and 104 video sequences from 52 individuals taken by the Kinect. It contains nine different facial variations (including different facial expressions, illuminations, occlusions, and poses) within two separate sessions. We conducted benchmark evaluations on the proposed database using a number of baseline face recognition methods. Specifically, we report the results of Eigenface [19], LBP [21], scale-invariant feature transform (SIFT) [33], and local Gabor binary patterns (LGBP) [34] for 2-D and 2.5-D-based face recognition, and applied both the rigid method (i.e., ICP) [35] and the nonrigid method (i.e., TPS) [36] for 3-D face recognition. Score-level fusion of RGB and depth data was also conducted, which demonstrated significantly improved results using multimodal face recognition. The proposed KinectFaceDB is compared with the widely used FRGC database [24] with high-quality 3-D scans, which reveals the effect of different data qualities in 3-D face recognition. In addition, we also show in this paper that the proposed database can be applied to many other applications in addition to face recognition (such as facial demographic analysis and 3-D face modeling). The main contributions of this paper can be summarized as follows.

1) A complete multimodal (i.e., well-aligned and processed 2-D, 2.5-D, 3-D, and video-based face data) face database-based on the Kinect is built and thoroughly described.
2) Extensive experiments are conducted for the benchmark evaluations of 2-D, 2.5-D, and 3-D based recognition using standard face recognition methods.
3) Recognition results on both the KinectFaceDB and the FRGC are compared following the same protocol, which demonstrate the data quality differences between the Kinect sensor and the laser scanner in the context of face biometrics.
4) Three-dimensional face databases in the literature are reviewed and discussed.

The rest of this paper is structured as follows. In Section II, we review the 3-D face databases proposed in the literature. Details of the KinectFaceDB are described in Section III. In Section IV, experimental results of the benchmark evaluations and the proposed RGB-D fusion are provided. Comparative results of face recognition using both the KinectFaceDB and the FRGC are then reported in Section V. Finally we conclude in Section VI.

II. Review of 3-D Face Databases

This section gives an up-to-date review of 3-D face databases in the literature (a more complete list of generic face databases can be found in [37]). In biometrics, face databases serve as the standard platform for quantitative evaluation of different face recognition algorithms [38], which are essential for developing robust and reliable face recognition systems. Toward this goal, a large number of face databases have been proposed for two main purposes: 1) to test face recognition algorithms robust to one or multiple facial variations (e.g., the Yale face database B [39] includes faces with nine poses and 64 illumination variations) and 2) to assist with the development of face recognition algorithms using a specific data modality (e.g., the Honda/UCSD video face database [40], [41] was proposed for video-based face recognition). A relatively small number of face databases provide abundant types of facial variations and multiple data modalities, which can be used in different evaluation tasks (such as the FRGC database [24]). The most popular and widely used face databases in the literature include: FRGC database, FERET database [22], Alex and Robert (AR) database [42], pose-illumination-expression (PIE) database [43], and Olivetti Research Lab (ORL) database [44]. More recently, the labeled face in wild (LFW) [45] database was proposed to study face recognition in unconstrained scenarios, and the mobile biometry (MOBIO) database [46] was published to evaluate face/speaker recognition tasks in mobile environments.

In comparison to the large number of 2-D face databases, the number of available 3-D face databases is relatively small. Table I gives an overview of the published 3-D face databases with different statistics. In the table, it is clear that most of the existing databases (FRGC [24], ND-2006 [47], GavabDB [48], BJUT-3-D [49], and UMD-DB [50]) adopt high-quality laser scanners for data acquisition. Three-dimensional faces in BU-3-DFE [49], XM2VTSDB [51], and Texas 3-DFRD [52] are captured by using high-quality stereo imaging systems, which can yield similar data accuracy in comparison to the data obtained by laser scanners. Although those high-quality scanning systems can provide accurate facial details (e.g., wrinkles and eyelids) for analysis, their acquisitions are slow and require careful user cooperation. On the other hand, only the 3-D-RMA [53] database is captured by a low-quality 3-D inference scheme, where 4000 points of each identity are sampled by structured lights. The scanning scheme is similar to the one used in the Kinect. However, the resolution is much lower than the Kinect images (e.g., in KinectFaceDB, each cropped face consists in 65 536 points) and no texture

TABLE I
SUMMARY OF OFF-THE-SHELF 3-D FACE DATABASES

Database	Year of Publication	DB Size	No. of Subjects	3D Sensor	2D Texture	Expression	Illumination	Occlusion	Pose	Video
FRGC (Ver. 2.0) [24]	2005	121589	466	Minolta Vivid 900/910 [29]	✓	✓	✓			
ND-2006 [47]	2006	13450	888	Minolta Vivid 910	✓	✓				
GavabDB [48]	2004	549	61	Minolta VI-700 digitizer	✓	✓			✓	
3D-RMA [53]	1998	360	120	Structured Light					✓	
XM2VTSDB [51]	1999	1180	295	Stereo Camera	✓					✓
U-York [54]	n.a.	5250	350	n.a.	✓	✓	✓		✓	
BU-3DFE [55]	2006	2500	100	3DMD digitizer [56]	✓	✓			✓	
BJUT-3D [49]	2005	n.a.	500	CyberWare3030RGB/PS [57]	✓	✓				
Texas 3DFRD [52]	2010	1149	118	MU-2 stereo imaging system [58]	✓	✓				
UMB-DB [50]	2011	1473	143	Minolta Vivid 900	✓	✓		✓		

mapping is provided. Details of the sensor used in U-York database are unavailable.

A face database should include sufficient variations to test the robustness of recognition algorithms in different conditions. Facial variations such as expression, illumination, and pose are widely used in 2-D face databases. However, because 3-D data provides different information in comparison to 2-D data, 3-D face recognition methods differ in their treatment of facial variations. For example, it is well known that 3-D face recognition algorithms (e.g., [35] and [59]) are less affected by the illumination changes than 2-D methods; however, facial expression is still a major challenge in 3-D face recognition [60], [61]. For this reason, almost all 3-D face databases in our survey include facial expression variation (except for the 3-D-RMA database). The second commonly used variation in 3-D face databases is the pose variation. It can demonstrate the advantage of using 3-D by aligning two face meshes via rigid transformations (e.g., using the iterative closest point (ICP) method [27], [28]), which is considered as a more difficult problem in 2-D face recognition. In the FRGC database and the U-York database, illumination variation is also included. In addition to the commonly considered variations, partial occlusion is a very challenging problem for both 2-D and 3-D face recognition but only few databases contain this variation [42], [62], [63]. Recently, the UMB-DB [63] is proposed for the evaluation of 3-D face recognition methods under occluded conditions.

Almost all 3-D face databases provide multiple modalities (i.e., both the 2-D and 3-D images, except for the 3-D-RMA database). For the databases based on stereo imaging techniques (i.e., BU-3-DFE, XM2VTSDB, and Texas 3-DFRD), their 3-D images are inferred from two or more RGB cameras, thus the 2-D/3-D mapping is achieved directly. However, for the databases based on laser scanners (i.e., FRGC, ND-2006, GavabDB, BJUT-3-D, and UMD-DB), the 2-D texture images are taken by an external RGB camera. Therefore the 2-D image and the 3-D image of the same face are originally not aligned. The alignment can be achieved using additional facial landmarks and warping algorithms such as the thin plate spline (TPS) method [64]. It should be noted that only one database (XM2VTSDB) in Table I provides also the video data, whose video sequences are captured by the RGB camera. Three-dimensional video sequences are not provided in any of the existing 3-D face databases we reviewed. This is because traditional 3-D scanners are unable to capture 3-D data in real time. Unfortunately, the lack of 3-D video data limits 3-D face recognition methods in the literature to still 3-D images, whereas 3-D video sequences can potentially provide complementary information to discriminate different identities.

III. KINECT FACE DATABASE

In this paper, we introduce a multimodal face database (with well-aligned 2-D, 2.5-D, 3-D, and video data) based on the Kinect sensor which can be a valuable addition to the repository of existing 3-D face databases. We will first introduce the structure and the acquisition environment of the proposed KinectFaceDB in Sections III-A and B, respectively. Then the details of the imaging method from multiple modalities using the Kinect as well as the RGB-D alignment procedure are described in Section III-C. The postprocessing steps (including noise removal and facial landmarking) are then presented in Section III-D. Finally, we discuss the potential usages of the proposed KinectFaceDB in Section III-E.

A. Database Structure

Fifty-two volunteers participated in the database recording, with 38 males and 14 females. The participants were born between 1974 and 1987, and are from different countries with different ethnicity. We categorize their ethnicity into the following classes (with the number of participants in each class): Caucasian (21), Middle East/Maghreb (11), East Asian (10), Indian (4), African-American (3), and Hispanic (3). The participants were asked to attend two different sessions. There are 5–14 day intervals between the two sessions, where the same recording protocol is applied. A meta-data file including the information of gender/birth year/ethnicity/with or without glasses/capturing time/session is associated with each identity. The demographic classification of the proposed KinectFaceDB can be seen in Fig. 1.

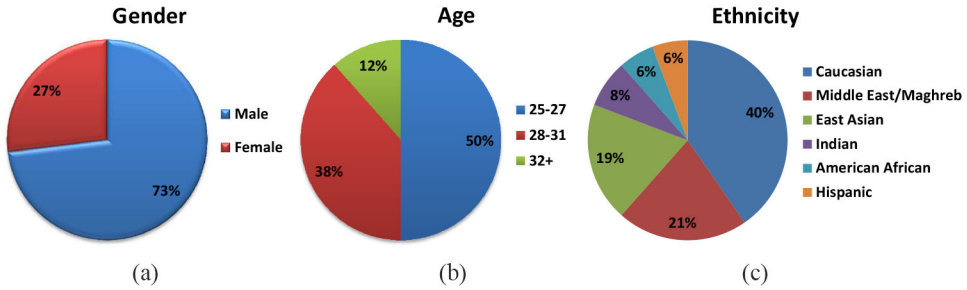


Fig. 1. Demographic partition of the proposed KinectFaceDB by (a) gender, (b) age, and (c) ethnicity.

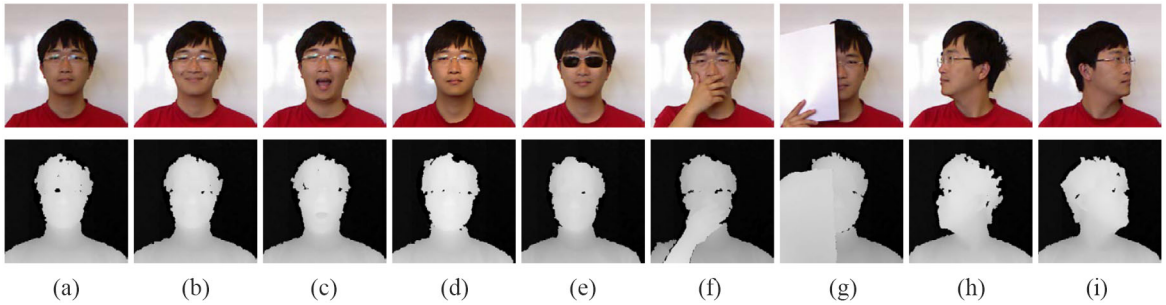


Fig. 2. Illustration of different facial variations captured in our database. (a) Neutral face. (b) Smiling. (c) Mouth open. (d) Strong illumination. (e) Occlusion by sunglasses. (f) Occlusion by hand. (g) Occlusion by paper. (h) Right face profile. (i) Left face profile. Upper: the RGB images. Lower: the depth maps aligned with above RGB images.

In each session, four types of data modalities are captured for each identity: 1) the 2-D RGB image; 2) the 2.5-D depth map; 3) the 3-D point cloud; and 4) the RGB-D video sequence. We carefully designed nine facial variations in both sessions: i.e., neutral face, smiling, mouth open, strong illumination, occlusion by sunglasses, occlusion by hand, occlusion by paper, right face profile, and left face profile. Examples of different variations are illustrated in Fig. 2. All images were taken under controlled conditions, but no restraints on clothing (clothes, glasses, etc.), make-up, or hair style were imposed on the participants.

We also devised a protocol to record the RGB-D video sequences for each person in the two sessions. The protocol consists of slow head movements in both the horizontal (yaw) and the vertical (pitch) directions. Fig. 3 illustrates the procedure of recording one participant. The video sequence allows extraction of multiple frames with different poses (in addition to the left/right profile recorded in the still images) which can be used to test 2-D/3-D face recognition algorithms robust to pose. More importantly, video-based face recognition [65]–[68] can be studied on this dataset. In particular, the proposed KinectFaceDB can be used to develop face recognition methods based on 3-D video [30], [69], which is a new research topic in video-based face recognition. Moreover, accurate 3-D face models can be reconstructed from such video data via 3-D accumulation and refining [70].

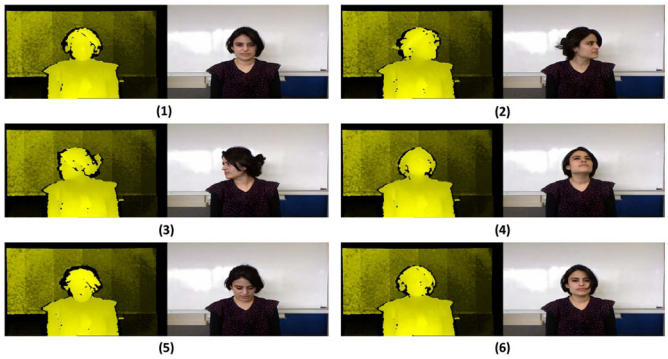


Fig. 3. Proposed procedure to record a video sequence: the head movement of a participant follows the path (1 $\rightarrow$ 2 $\rightarrow$ 3 $\rightarrow$ 4 $\rightarrow$ 5 $\rightarrow$ 6). The head yaw is first performed which follows the procedure (approximately): $0^\circ \rightarrow +90^\circ \rightarrow -90^\circ \rightarrow 0^\circ$; then the head pitch is performed as: $0^\circ \rightarrow +45^\circ \rightarrow -45^\circ \rightarrow 0^\circ$.

B. Acquisition Environment

We set up a controlled indoor environment (natural light at daytime, plus moderate indoor LED diffusion light) for the database recording. A Kinect is mounted and stabilized (that is in parallel to the ground by adjusting its tilt) on top of a laptop. The participants[2] were asked to sit in front of the Kinect sensor at a distance (ranging from 0.7 m to 0.9 m) and to follow the predefined acquisition protocol (which conforms to the database structure described in Section III-A). A white board is placed behind each participant with fixed distance to the Kinect at 1.25 m, in order to produce a simple background which can be easily filtered. A LED lamp is set in front of the participant to yield the illumination variation. Different sunglasses and a piece of paper are used to produce the occlusion variation. A human operator is required to sit in front of the laptop (in the opposite position to the participant) in order to monitor and control the acquisition process.

[2]Fifty-two Ph.D. students from EURECOM: http://www.eurecom.fr/

A software application (based on the OpenNI [71] library) is developed for the database recording. The software automatically captures, processes, and organizes faces of the participants in accordance with the predefined database structure. Since we focus on the facial region, we crop the captured RGB image and the depth image using a predefined ROI (with the size of 256×256). The cropping scheme ensures that the captured faces have a simple/uniform background (the white board only, and thus easy to segment). Also, it minimizes the differences between the RGB image and depth map after alignment (we will give details in Section III-C3). The imaging method of this acquisition process is presented in the next section.

C. Acquisition Process

The Kinect sensor was primarily designed for entertainment purposes and integrates different sensing/displaying/processing functionality (based on different data modalities ranging from video/audio to depth). Since our goal focuses on RGB and depth data, in this section, we summarize the 3-D imaging procedure of the Kinect, and show the alignment process for RGB and depth images, which yields the final output (still RGB-D images, 3-D point cloud, and video sequences) in our database.

1) RGB and Depth Imaging From the Kinect: The Kinect sensor contains three main components for the RGB-D sensing: an infrared (IR) laser emitter, a RGB camera, and an IR camera. The RGB camera captures the RGB image I_{RGB} directly, whereas the laser emitter and the IR camera act together as an active depth sensor to retrieve the distance information from the scene.

Freedman *et al*. [72] introduced a triangulation process for the depth measurement of the Kinect based on the IR laser emitter and the IR camera. In the proposed system, a predesigned pattern of spots (created by transillumination through a raster) are projected to the scene by the IR laser emitter, and the reflection of the pattern is captured by the IR camera. The captured pattern is then compared with a reference pattern (at the predefine plane with a known distance) in order to produce a disparity map $I_{Disparity}$ with the disparity value d at each point. From the obtained disparity map $I_{Disparity}$ it is straightforward to deduce the depth map I_{Depth} via a simple triangulation method. A simplified version to describe the triangulation of the Kinect is suggested in [7]

$$z_{\text{world}}^{-1} = \left(\frac{m}{f \times b}\right) \times d' + \left(Z_0^{-1} + \frac{n}{f \times b}\right) \tag{1}$$

where z_{world} is the distance between the Kinect and the real-world location (namely the depth, in the unit of *mm*); d' is the normalized disparity value by normalizing the raw disparity value d between 0 and 2047, thus $d = m * d' + n$ where m and n are the de-normalization parameters; b and f are the base length and focal length, respectively; and Z_0 is the distance between the Kinect and the predefined reference pattern. The calibration parameters including b, f, and Z_0 are estimated and provided by the device vendor. For more details of the model described in (1), please refer to [7].

With the triangulation process, the Kinect outputs a RGB image I_{RGB} [where $I_{RGB}(x, y) = \{v_R, v_G, v_B\}$, and v_R, v_G, v_B refer to the values of R, G, B channels at image location (x, y)] and a depth map I_{Depth} [where $I_{Depth}(x, y) = z_{world}$, and z_{world} indicates the depth value at image location (x, y)] simultaneously, with the image size of 640×480. Based on the obtained RGB and depth images, 3-D point coordinates can be computed and aligned RGB-D face data can be obtained using the following methods.

2) Converting to 3-D Face Data: For 3-D face recognition, 3-D coordinates (with respect to a predefined reference origin) for different faces need to be computed. Thanks to the design of the Kinect, the 3-D coordinates calculation is straightforward. Given the depth map $I_{Depth}(x, y) = z_{world}$ obtained in Section III-C, the 3-D coordinates of each point $(x_{\text{world}}, y_{\text{world}}, z_{\text{world}})$ can be calculated from its image location (x, y) as below

$$x_{\text{world}} = -\frac{z_{\text{world}}}{f}(x - x_0 + \delta x) \tag{2}$$

$$y_{\text{world}} = -\frac{z_{\text{world}}}{f}(y - y_0 + \delta y) \tag{3}$$

where (x_0, y_0) is the principal location of the depth image, and δx and δy represent the corrections due to the lens distortions, where δx and δy are estimated in advance and provided by the device vendor.

Based on the above projection, we can compute the 3-D coordinates of each precropped face depth image (the precropping scheme is described in Section III-B) and store them in the 3-D format in KinectFaceDB. Then the 3-D data can be used in the evaluation of 3-D face recognition methods and data quality assessment in the experiment section.

3) RGB-D Alignment for Face Data: For face recognition based on both the RGB and depth images, establishing the correspondences between the RGB values and the depth/3-D values at the same location on a face is an essential step. For example, designing a RGB-D descriptor which can summarize features from both the RGB and depth values jointly for each pixel might potentially reveal important facial characteristics. However, due to the intrinsic architecture of the Kinect sensor (where the RGB and depth images are sampled separately from two different cameras with a displacement between them), the RGB image, and the depth map captured by the Kinect are not well aligned. Therefore, we further project the depth value from the IR camera plane to the RGB camera plane. From the depth map, we have already estimated the corresponding 3-D coordinate $(x_{\text{world}}, y_{\text{world}}, z_{\text{world}})$ of each pixel using the method presented in Section III-C2, then the projection from the 3-D coordinates to the 2-D RGB camera plane can be achieved by using the traditional Tsai's camera model [73]. First, the 3-D coordinates based on the IR camera are transformed to the 3-D coordinate system defined by the RGB camera using affine transformation

$$\begin{bmatrix} x_{\text{world}}' \\ y_{\text{world}}' \\ z_{\text{world}}' \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ \vec{0} & 1 \end{bmatrix} \begin{bmatrix} x_{\text{world}} \\ y_{\text{world}} \\ z_{\text{world}} \\ 1 \end{bmatrix} \tag{4}$$

where $R \in \mathbb{R}^{3\times3}$ is the rotation matrix and $T \in \mathbb{R}^{3\times1}$ is the translation vector. Then the 3-D coordinates based on the RGB

Fig. 4. Illustration of the RGB-D alignment: the depth image (top-left) is aligned with the RGB image (top-right) captured by the Kinect at the same time; the bottom image shows the alignment effects by overlapping the aligned depth image and the RGB image.

camera can be mapped to the ideally undistorted RGB camera plane as

$$\begin{bmatrix} x_{RGB} \\ y_{RGB} \\ 1 \end{bmatrix} = \frac{f_{RGB}}{z_{world'}} \begin{bmatrix} x_{\text{world}'} \\ y_{\text{world}'} \\ z_{\text{world}'} \end{bmatrix} \tag{5}$$

based on the focal length f_{RGB} of the RGB camera. Finally, the true location $(x_{RGB'}, y_{RGB'})$ of a 3-D point in the RGB camera plane is recovered by correcting the distortions and mapping to the RGB image origin

$$\begin{bmatrix} x_{RGB'} \\ y_{RGB'} \\ 1 \end{bmatrix} = VD \begin{bmatrix} x_{RGB} \\ y_{RGB} \\ 1 \end{bmatrix} \tag{6}$$

where $D \in \mathbb{R}^{3\times3}$ and $V \in \mathbb{R}^{3\times3}$ are estimated to correct the uncertainties due to imperfections in hardware timing and digitization, as well as the lens distortions. Some previous works made additional efforts to calibrate the intrinsic and extrinsic parameters of the RGB camera based on an extra RGB camera [31] or the embedded IR camera [32]. In our method, we directly adopt the Kinect factory calibration parameters which can also produce satisfactory alignment results. Illustration of the RGB-D alignment is shown in Fig. 4. In the figure, one can observe the geometrical distortions (in addition to the translation at the image boundary) when remapping the depth map to the RGB image.

Because we apply a precropping scheme (as described in Section III-B) to the mapped depth image, the large information loss and distortions at the image boundary are not included in our final output. We then store the RGB image and the aligned depth map (using the original scale in *mm*), respectively. Once we found the correspondences between the RGB image and the depth image, it is straightforward to map the RGB color to the 3-D points. The 3-D point cloud with corresponding color mappings are then recorded. Visualization of a 3-D face after color mapping is displayed in Fig. 5 (with background removal using a threshold τ). Finally, we store the video sequences of the aligned RGB and depth frames from both the RGB camera and the IR camera using the protocol described in Section III-A. The video-based face data can then be used for multiple purposes ranging from the video-based face recognition using either 2-D [65]–[68] or 3-D [30], [69], to the dense face model reconstruction [70] from 3-D frames.

Fig. 5. Illustration of the color mapping of a 3-D face from left to right views.

D. PostProcessing

We include two types of postprocessing after data acquisition: namely the noise removal and the facial landmarking. Faces extracted after postprocessing are more appropriate for face recognition.

1) Noise Removal: Unlike the RGB images, the depth images captured by the Kinect are relatively noisy and inaccurate. A notable problem is that the depth values on some pixels can be sensed as 0 mm when their true values are not zero. This may occur for the several reasons: 1) the point is too far (out of the sensing range); 2) the point is too close (which lies in a blind region due to the limited field of view for the projector and the IR camera); 3) the point is in shadow cast by the projector (where the IR light cannot reach); or 4) the surface reflects poor IR light (such as hairs or specular surfaces) [2]. Because our acquisition environment is well controlled (in which the face is captured at a moderate distance, with the uniform background), the sensing noise caused by 1) and 2) are automatically eliminated. The sensing noise caused by 3) and 4) could be removed by setting a nonnegative threshold and the missing values can be filled in various manners (e.g., median/mean filters, morphological operations, Markov random fields). The method applied in our evaluation is described in Section IV-B.

Another noteworthy issue is the depth resolution. In comparison to traditional 3-D scanners (such as the KONICA Minolta scanner, which has the depth resolution at 0.1 mm within the sensing range, and was used to build the FRGC database [24]), depth resolution of the Kinect is much lower (around 2 mm at the distance of 1 m [7]). Therefore, details on a face (for example wrinkles and eyelids) cannot be recorded using the Kinect. Such facial details can be used as the soft biometric characteristics [74], [75], which can also provide useful information for the identity discrimination. Although the regions with large depth transitions (such as the nose region, which are less affected by the resolution problem) could contribute more in face recognition, it is unclear how much the lower resolution can decrease the recognition performance. In this paper, we evaluated the resolution differences using a number of standard 2.5-D/3-D face recognition algorithms on both the FRGC database and the proposed KinectFaceDB following the same protocol and report the quantitative results for comparison in Section V. Fig. 6 gives a visual illustration of

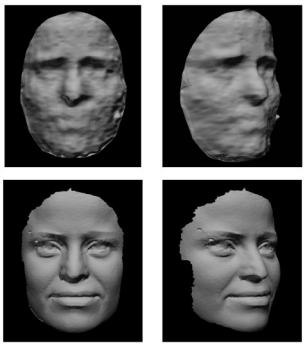


Fig. 6. Cropped and smoothed 3-D faces captured by the Kinect (upper row) and Minolta scanner (lower row) of the same person, with the frontal view (left column) and the side view (right column). It is clear that the 3-D face from Minolta contains more details (wrinkles, eyelids, etc.).

Fig. 7. Six anchor points on a face. 1: Left eye center. 2: Right eye center. 3: Nose-tip. 4: Left mouth corner. 5: Right mouth corner. 6: Chin.

the resolution difference between the Kinect and the Minolta laser scanner, where the two 3-D faces from the same person are captured using different sensors.

2) Facial Landmarking: In order to perform facial region extraction and normalization in face recognition, we define six anchor points on the face (namely the left eye center, the right eye center, the nose-tip, the left mouth corner, the right mouth corner, and the chin, as shown in Fig. 7). The anchor points are first manually annotated on the RGB image. Then the corresponding locations on the depth map and the 3-D points can be directly found based on the pointwise correspondences we established. Note that the left/right face profiles are not annotated because only one side of the face is available. Even if the occluded faces (faces occluded by sunglasses, hand, and paper) are not fully visible, we provide the full annotations on those occluded faces by estimating the anchor points within the occluded region (similar to the annotation method used in the AR face database [42]).

E. Potential Database Usages in Addition to Face Recognition

The primary application of the proposed KinectFaceDB is for face recognition. However, it can also be applied to other research tasks. In this section, we suggest the usage of the KinectFaceDB for two possible applications which have recently received a large amount of attention from computer vision researchers.

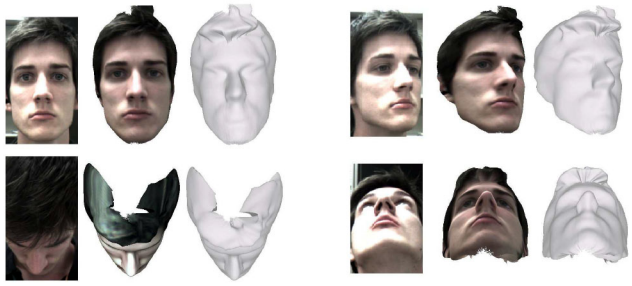


Fig. 8. Example of a 3-D face model generated using a video of cooperative head movement (images taken from [70]).

1) Facial Demographic Analysis: One potential application is the facial demographic analysis, including gender recognition, age estimation, and ethnicity identification based on the face images. Since the population of KinectFaceDB is consisted of diverse categories (as shown in Fig. 1), it is appropriate for the demographic analysis task based on the Kinect. According to a recent survey [76], most of the literature works for gender recognition and ethnicity identification are based on the FERET face database using 2-D images. Nevertheless, Lu *et al*. [77] suggested that the 3-D shape of human face can provide more discriminative information on demographic classification in comparison to the 2-D information. As a device which can capture both 2-D textures and 3-D shapes, the Kinect can act as an ideal sensor for the task of multimodal-based facial demographic analysis.

2) Three-Dimensional Face Modeling: Recently, dense 3-D modeling using the Kinect has attracted vast amount of attention from the computer vision and computer graphic researchers. Pioneering works [4], [5] have demonstrated how to build a dense 3-D map of indoor scene/object by camera tracking using sparse features (e.g., SIFT [33] and FAST [78]) and optimize the 3-D points aggregation, taking advantage of the real-time, low-cost, and ease of RGB-D mapping from the Kinect sensor. More recently, 3-D face modeling using the Kinect is introduced in [70], [79] to generate a 3-D avatar for video conferences and massive multiplayer online games (MMOGs). The proposed KinectFaceDB can also be used for dense 3-D face modeling. To this end, we recorded the video data following a similar protocol as described in [70]. An example of the 3-D model generated from such video data is illustrated in Fig. 8.

It can be observed in Fig. 8 that the generated 3-D face model from a video sequence has a much higher data quality than the 3-D face model obtained from a single depth image of the Kinect sensor (an example of single-shot based 3-D model can be found in Fig. 6). The video-based face modeling shown in Fig. 8 aggregates and averages data points from multiple single depth frames in a cylindrical coordinates system, so as to capture the complementary information brought by the given video sequence. Then a bilateral smoothing [80] is applied to remove noise while keeping edges. The generated 3-D faces from the Kinect-based video sequences have demonstrated comparable accuracy to laser scanned 3-D faces [70]. Nevertheless, the averaging strategy used here might not be

the optimal solution for 3-D face modeling. More advanced techniques might be applied to yield potentially improved accuracy. The videos recorded in our database can therefore serve as a standard dataset to evaluate the quality of 3-D face models generated by different algorithms.

In addition to the applications we discussed in this section (i.e., the facial demographic analysis and the 3-D face modeling), the KinectFaceDB may also be used in other applications, including but not limited to to: head pose estimation, facial expression analysis, 3-D face registration, RGB-D feature extraction, occlusion detection, and plastic surgery.

IV. Benchmark Evaluations

In this section, we report the results of benchmark evaluation on the proposed KinectFaceDB using standard face recognition methods. The benchmark evaluation is conducted for 2-D, 2.5-D, and 3-D based face recognition. For 2-D and 2.5-D based face recognition, PCA [19], LBP [21], SIFT [81], and LGBP [34]-based methods are tested. For 3-D based face recognition, both the rigid method based on ICP [35] and the nonrigid method based on TPS [36] are evaluated. In order to show that the integration of the RGB data and the depth data can improve the recognition results, we also provide the fusion results from both the RGB and depth images using score-level fusion. Details and results of the benchmark evaluation are provided in the following sections.

A. Baseline Techniques and Configurations

PCA [19] (i.e., the Eigenface method), LBP [21], SIFT [81], and LGBP [34]-based methods are selected as the baseline techniques for the 2-D and 2.5-D based face recognition. For 2-D face recognition, the methods are applied to the RGB images; whereas in 2.5-D face recognition, the depth images are used. In PCA-based method, the training data are used to build a low-dimensional face subspace, in which face recognition is conducted. For LBP-based method, the operator $LBP^{u2}_{8,2}$ is used to summarize features on 8×8 blocks. SIFT-based method extracts the key points from all training and testing images, where the similarity measure is achieved by key points matching. In LGBP-based method, the Gabor features are firstly computed, and then the LBP operator is applied to extract features on the Gabor filtered images. The fusion of RGB and depth images are achieved by directly extending the baseline techniques to use multiple modalities via score-level fusion. The nearest neighbor (NN) classifier is adopted for the identification task.

Two representative 3-D face alignment techniques (i.e., the ICP-based rigid registration [27], [28] and the TPS-based nonrigid registration [64]) are used in 3-D face recognition. In the ICP-based method [35], the volume difference based on Euclidean distance between two 3-D meshes are computed as the dissimilarity measure in face recognition after the rigid alignment by ICP. For TPS-based method, we tested the method proposed in [36] (which is the baseline method in the European project Tabula Rasa (EU FP7) [82]), that computes the warping parameters (WPs) based on the nonrigid TPS alignment after the rigid ICP registration. The face representation in the TPS-based method can thus be regarded as the deviations of an input face from the canonical face. Distances between the WPs of the probe face and the precalculated WPs of the gallery faces are computed for identification/verification by taking the mean cosine distance.

B. Preprocessing

Because the RGB data and the depth data in KinectFaceDB have already been aligned, given the facial landmarks, face cropping, and normalization can be directly achieved. Using the eye coordinates, we cropped, normalized, and downsampled the 2-D and 2.5-D faces into 96×96 dimensions.

The 3-D face cropping is achieved by preserving the vertices in a sphere with the radius of $100mm$, which are centered at $20mm$ away from the nose tip in the $+z$ direction. Afterwards, spikes are removed by thresholding and a hole filling procedure is applied (the holes and spikes are interpolated linearly to form a complete mesh, the values to fill holes and spikes are estimated by taking the mean of valid pixels in the neighborhood of a 5×5 patch). Finally, a bilateral smoothing filter [80] is employed to remove the white noise while preserving the edges (the example of preprocessed 3-D face from the Kinect can be found in Fig. 6).

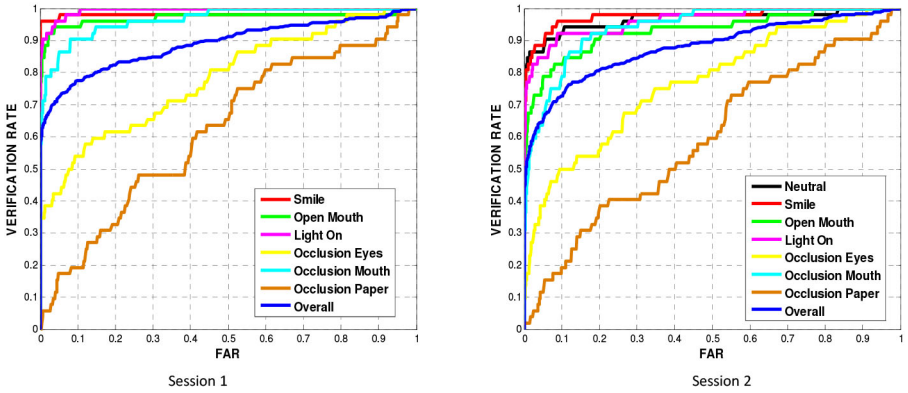
C. Evaluation Protocol

We conduct both identification and verification in the benchmark evaluation. In both modes, we use the neutral faces from session 1 as the gallery face. Recognition results of each variation (except for the left/right profiles, since sophisticated face alignment is needed in both the 2-D and 2.5-D-based recognition under large pose variations) from both sessions 1 and 2 are reported. Then the overall identification/verification rates are reported for all tested facial variations in both sessions. In our evaluation, the rank-1 identification rate and the verification rate where the false acceptance rate (FAR) equals to 0.001 are reported.

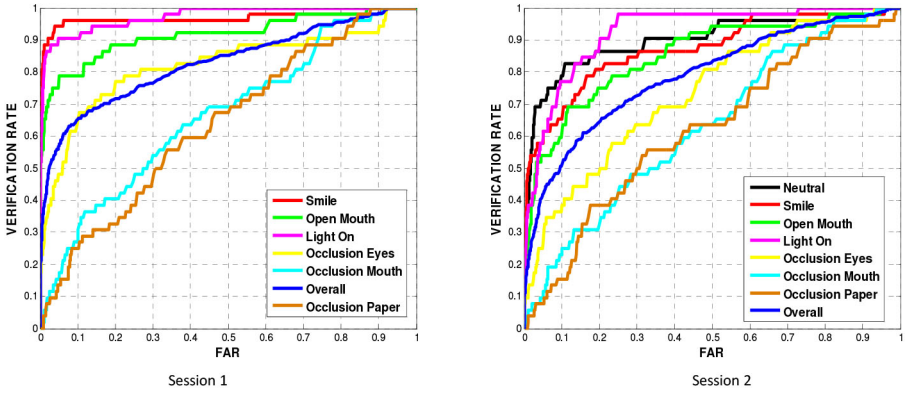
The receiver operating characteristic (ROC) curves of the PCA-based method for both 2-D and 2.5-D-based face recognition are shown in Fig. 9. In the figure, it is clear that different facial variations lead to different verification results, and the time-elapsing between the two sessions also significantly affect the recognition performance.

D. Evaluation Results

1) Results of 2-D Face Recognition: Tables II and III show the identification rates and the verification rates of 2-D face recognition methods based on PCA, LBP, SIFT, and LGBP under different facial variations for the two sessions. As can be observed in the table, PCA is not robust to large local distortions (such as extreme facial expressions and partial occlusions), since the local distortions can alter the entire face representation in the PCA space. In the contrary, since LBP, SIFT, and LGBP are local-based methods, they are more robust to such local distortions. In the table, it is clear that the partial occlusions (i.e., sunglasses, hand on face, and paper on



(a)



(b)

Fig. 9. ROC curves of 2-D and 2.5-D-based face recognition using PCA [19] for different facial variations. (a) ROC curves of 2-D face verification using PCA. (b) ROC curves of 2.5-D face verification using PCA.

TABLE II
RANK-1 IDENTIFICATION RATE FOR 2-D FACE RECOGNITION

Session	Method	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	**Overall**
Session 1	PCA	N/A	96.15%	78.85%	90.38%	38.46%	73.08%	7.69%	64.10%
	LBP	N/A	**100%**	96.15%	**100%**	90.38%	**100%**	67.31%	92.31%
	SIFT	N/A	**100%**	96.15%	88.46%	84.62%	94.23%	57.69%	86.86%
	LGBP	N/A	**100%**	**98.08%**	98.08%	**92.31%**	98.08%	**78.85%**	**94.23%**
Session 2	PCA	82.69%	78.85%	67.31%	73.08%	19.23%	51.92%	1.92%	53.57%
	LBP	**100%**	98.08%	**94.23%**	**100%**	**92.31%**	94.23%	57.69%	90.93%
	SIFT	98.08%	98.08%	86.54%	78.85%	57.69%	82.69%	17.31%	74.18%
	LGBP	**100%**	**100%**	92.31%	**100%**	88.46%	**100%**	**84.62%**	**95.05%**

face) are very challenging to all tested methods, especially in the verification mode. Nevertheless, LGBP-based method is more robust to partial occlusions in comparison to the other methods, and gives the best overall performances. This result conforms to the findings in [34], where the authors suggested that the multiresolution/multiorientation-based Gabor decomposition and the local patch-based representation of LGBP can enhance its robustness to partial occlusions.

2) Results of 2.5-D Face Recognition: Tables IV and V illustrate the evaluation results on 2.5-D face data. Although previous studies (according to [59]) suggested to directly apply PCA on the 2.5-D range images, results of PCA-based method in our experiment is not as good as the ones obtained with LBP and LGBP, especially for large facial variations. On the other hand, LBP yields the best overall identification/verification results on the depth images in both sessions (even if LBP was primarily designed as a texture description). Unlike the results for 2-D face recognition, the Gabor features used in LGBP cannot improve the recognition results based on LBP. It should be noted that SIFT cannot yield meaningful results on the depth images, since it identifies key points at salient image gradients. Because the depth map is highly smooth, the SIFT-based method is inappropriate for 2.5-D face recognition.

As observed in the results, the LBP-based method (as well as its depth-specific variants such as [83], [84]) is more appropriate to represent and discriminate depth face patterns. It should be noted that the 2.5-D depth maps result in lower identification/verification rates in comparison to the 2-D intensity

TABLE III
VERIFICATION RATE ($FAR = 0.001$) FOR 2-D FACE RECOGNITION

Session	Method	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	**Overall**
Session 1	PCA	N/A	96.15%	76.92%	84.62%	34.62%	51.92%	0%	58.01%
	LBP	N/A	96.15%	90.38%	96.15%	**88.46%**	92.31%	40.38%	**75.96%**
	SIFT	N/A	**100%**	88.46%	73.08%	61.54%	78.85%	5.77%	43.59%
	LGBP	N/A	**100%**	**92.31%**	**98.08%**	80.77%	**94.23%**	**57.69%**	71.79%
Session 2	PCA	73.08%	61.54%	51.92%	55.77%	7.69%	26.92%	1.92%	40.11%
	LBP	92.31%	82.69%	73.08%	88.46%	65.38%	67.31%	17.31%	59.34%
	SIFT	90.38%	84.62%	48.08%	57.69%	30.77%	59.62%	0%	38.74%
	LGBP	**96.15%**	**94.23%**	**84.62%**	**96.15%**	**67.31%**	**82.69%**	**42.31%**	**61.26%**

TABLE IV
RANK-1 IDENTIFICATION RATE FOR 2.5-D FACE RECOGNITION

Session	Method	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	**Overall**
Session 1	PCA	N/A	84.62%	57.69%	76.92%	36.54%	7.69%	5.77%	44.87%
	LBP	N/A	**94.23%**	**84.62%**	**96.15%**	**84.62%**	65.38%	19.23%	**74.04%**
	SIFT	N/A	7.84%	3.92%	7.84%	0%	1.96%	1.96%	1.63%
	LGBP	N/A	90.38%	82.69%	94.23%	59.62%	**69.23%**	**44.23%**	73.40%
Session 2	PCA	46.15%	42.31%	36.54%	30.77%	13.46%	5.77%	0%	25%
	LBP	**92.31%**	73.08%	**80.77%**	**94.23%**	**73.08%**	**38.46%**	5.77%	**65.38%**
	SIFT	5.88%	1.96%	0%	5.88%	3.92%	0%	3.92%	1.12%
	LGBP	80.77%	**75%**	65.38%	78.85%	34.62%	**38.46%**	**25%**	56.87%

TABLE V
VERIFICATION RATE ($FAR = 0.001$) FOR 2.5-D FACE RECOGNITION

Session	Method	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	**Overall**
Session 1	PCA	N/A	67.31%	38.46%	48.08%	15.38%	0%	0%	17.95%
	LBP	N/A	75%	**75%**	**78.85%**	**34.62%**	19.23%	1.92%	**41.03%**
	SIFT	N/A	1.96%	0%	0%	0%	0%	0%	0.33%
	LGBP	N/A	**76.92%**	69.23%	57.69%	28.85%	**36.54%**	**3.85%**	36.86%
Session 2	PCA	21.15%	15.38%	15.38%	17.31%	7.69%	0%	0%	8.52%
	LBP	**55.77%**	34.62%	15.38%	**34.62%**	**23.08%**	**11.54%**	**5.77%**	**26.92%**
	SIFT	0%	0%	0%	0%	0%	0%	0%	0%
	LGBP	46.15%	**50%**	**23.08%**	**34.62%**	7.69%	**11.54%**	1.92%	26.37%

images (for all tested methods). This is because the depth scanning quality of the Kinect is relatively low. Nevertheless, the depth map provides complementary information to the intensity image, and therefore 2.5-D-based method can be integrated with 2-D-based method to achieve multimodal face recognition, which can yield higher recognition results. We demonstrate the combination of 2-D and 2.5-D information in Section IV-E using score-level fusion.

3) Results of 3-D Face Recognition: In addition to the direct usage of 2.5-D depth map, the use of 3-D surface registration algorithms (such as ICP and TPS) for 3-D faces is a popular approach for 3-D face recognition [25], [26]. In this paper, we tested two 3-D face recognition methods based on ICP and TPS on the proposed KinectFaceDB, respectively. Tables VI and VII show the identification and verification results for the ICP and TPS-based methods under different variations in both sessions. In the table, we can observe that 3-D face recognition methods used in this paper are more robust to certain type of variations (e.g., the time-elapsing for neutral faces) than the 2.5-D depth-based methods. However, it generates inferior results for facial expression and occlusion variations than the 2.5-D methods. The results indicate that the ICP and TPS-based methods can better align different faces than the 2.5-D methods, however, they are unable to handle large local facial distortions (such as facial expression). One possible solution is to use deformable face models such as the annotated face model (AFM) [60] and the active appearance model (AAM) [85], [86] to overcome the local distortion problem caused by facial expression for the Kinect images.

In the table, the TPS-based method generates better recognition results in most cases than the ICP-based method. This is because the nonlinear alignment acted in TPS can partially handle the facial expression problem to some extent (which conforms to the results in [36]). Nevertheless, both of the ICP and TPS-based methods cannot handle the partial occlusion problem. For faces occluded by hand and paper, the large surface distortions completely mislead the face registration (thus we do not report their recognition results). In addition, as shown in our experiment, the tested methods cannot yield reliable results for the verification task using low quality Kinect data.

TABLE VI
RANK-1 IDENTIFICATION RATE FOR 3-D FACE RECOGNITION

Session	Method	Neutral	Smile	Open mouth	Illumination	Sunglasses	**Overall**
Session 1	ICP	N/A	**63.46%**	26.92%	51.92%	**50%**	38.46%
	TPS	N/A	59.62%	**47.06%**	**71.15%**	38.46%	**44.53%**
Session 2	ICP	46.15%	42.31%	23.08%	44.23%	38.46%	32.37%
	TPS	**78.85%**	**46.15%**	**38.46%**	**67.31%**	**53.85%**	**48.70%**

TABLE VII
VERIFICATION RATE ($FAR = 0.001$) FOR 3-D FACE RECOGNITION RECOGNITION

Session	Method	Neutral	Smile	Open mouth	Illumination	Sunglasses	**Overall**
Session 1	ICP	N/A	**26.92%**	1.92%	26.92%	**13.46%**	**16.92%**
	TPS	N/A	19.23%	**23.53%**	**28.85%**	7.69%	12.11%
Session 2	ICP	23.08%	**25%**	1.92%	15.38%	**17.31%**	14.42%
	TPS	**32.69%**	17.31%	**15.38%**	**28.85%**	**17.31%**	**19.81%**

TABLE VIII
FUSION OF RGB AND DEPTH FOR FACE RECOGNITION RANK-1 IDENTIFICATION RATE

Session	Method	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	**Overall**
Session 1	PCA	N/A	96.15%	88.46%	100%	59.62%	78.85%	38.46%	76.92%
	LBP	N/A	**100%**	98.08%	100%	**92.31%**	**98.08%**	**94.23%**	**97.12%**
	LGBP	N/A	**100%**	**100%**	**100%**	90.38%	**98.08%**	92.31%	96.79%
Session 2	PCA	82.69%	82.69%	71.15%	90.38%	46.15	57.69%	30.77%	65.93%
	LBP	**100%**	**100%**	**98.08%**	98.08%	**94.23**	94.23%	**69.23%**	93.41%
	LGBP	**100%**	**100%**	96.15%	**100%**	88.46%	**98.08%**	**76.92%**	**94.23%**

4) General Remarks: In addition to the occlusions used in previous works (e.g., [42], [62], and [63]), we introduce a novel facial occlusion in the KinectFaceDB (namely paper on face). Comparing with the traditional partial occlusions (e.g., sunglasses and hand on face), the occlusion of paper on face is very challenging in face recognition according to the results we obtained. In this variation, only the left half of a face is visible. Although previous studies suggested that only half of a face [87] can provide sufficient information for face recognition (due to the face symmetry), the paper occlusion introduces large, nonuniform noise so that it needs to be deliberately handled. This controlled occlusion is similar to many occlusion cases encountered in crowded scenes. In addition, the facial asymmetry is also lost due to the paper occlusion, which can also provide useful clues for face recognition [88]. As a possible solution, as suggested in [89], explicit occlusion analysis could be useful to remove the features extracted from the occluded part (i.e., the paper), so as to improve the recognition results by only taking into account the nonoccluded facial part.

E. Fusion of RGB and Depth Face Data

To justify that the Kinect is more helpful than sole RGB-based cameras for face recognition, we conduct an additional fusion step to combine both the RGB (2-D) and the depth (2.5-D) face information from the Kinect in face recognition. The weighted sum fusion strategy is thus adopted for this purpose. First, z-score normalization is applied to all the dissimilarity scores of both matchers (of RGB and depth) for all the gallery and probe faces, respectively

$$\tilde{s}_{RGB}(i) = \frac{s_{RGB}(i) - \mu_{RGB}}{\delta_{RGB}} \tag{7}$$

$$\tilde{s}_{D}(i) = \frac{s_{D}(i) - \mu_{D}}{\delta_{D}} \tag{8}$$

where i is the face index, $\{\mu_{RGB}, \mu_D\}$ and $\{\delta_{RGB}, \delta_D\}$ are the means and standard deviations of all the dissimilarity scores from all gallery faces for the RGB image and the depth image, respectively. Then, for all the gallery faces and probe faces, their fused dissimilarity scores are computed as the weighted sum of their RGB scores and depth scores

$$\tilde{s}_{RGB-D}(i) = \frac{w_{RGB}\tilde{s}_{RGB}(i) + w_D\tilde{s}_D(i)}{w_{RGB} + w_D} \tag{9}$$

where i is the face index, w_{RGB} and w_D are the weights for the RGB scores and the depth scores, respectively. In our experiment, the combination weights are determined by grid search on all the gallery and probe faces. Finally, instead of using the original scores computed from RGB and depth separately, the fused dissimilarity scores are used for both identification and verification.

Tables VIII and IX illustrate the fusion results from both the RGB and depth using PCA, LBP, and LGBP (SIFT is not used because it cannot capture the correct information from depth images as shown in Section IV-D2). From the results, it is clear

TABLE IX
FUSION OF RGB AND DEPTH FOR FACE RECOGNITION VERIFICATION RATE ($FAR = 0.001$)

Session	Method	Neutral	Smile	Open mouth	Illumination	Sunglasses	Hand on face	Paper on face	**Overall**
Session 1	PCA	N/A	96.15%	84.62%	90.38%	30.77%	51.92%	0%	58.65%
	LBP	N/A	**100%**	92.31%	98.08%	**90.38%**	92.31%	15.38%	83.65%
	LGBP	N/A	**100%**	**94.23%**	**98.08%**	76.92%	**98.08%**	**55.77%**	**86.22%**
Session 2	PCA	75%	71.15%	46.15%	71.15%	5.77%	19.23%	1.92%	41.48%
	LBP	**96.15%**	80.77%	71.15%	92.31%	**63.46%**	69.23%	19.23%	73.63%
	LGBP	94.23%	**90.38%**	**88.46%**	**96.15%**	61.54%	**78.85%**	**46.15%**	**78.57%**

TABLE X
KINECTFACEDB VERSUS FRGC

Mode	DB	2.5D PCA	2.5D LBP	2.5D LGBP	3D ICP	3D TPS
Rank-1 Identification Rate	KinectFaceDB	46.15%	**92.31%**	80.77%	46.15%	78.85%
	FRGC	68.18%	93.94%	93.94%	58.08%	**96.97%**
Verification Rate ($FAR = 0.001$)	KinectFaceDB	21.15%	**55.77%**	46.15%	23.08%	32.69%
	FRGC	53.54%	81.82%	83.33%	39.90%	**87.37%**

that the fusion process significantly improves the results from sole RGB-based face recognition. For example, the overall rank-1 identification rate of session 1 is increased from 64.10% to 76.92%, 92.32% to 97.12%, and 94.23% to 96.79% for the PCA, LBP, and LGBP-based methods, respectively. This experiment demonstrates that 3-D information provided by the Kinect is helpful to improve the recognition performance using 2-D images.

V. DATA QUALITY ASSESSMENT OF KINECTFACEDB AND FRGC

It is straightforward to visually observe the 3-D data quality differences between the Kinect and a high-quality laser scanner (e.g., the Minolta, see Fig. 6). The device parameters such as the depth accuracy and the depth resolution also indicate the differences in terms of data quality between the two sensors. Recently, additional efforts have been made to better understand the accuracy of the Kinect [7]. However, it is not straightforward to quantitatively evaluate the data quality differences between the two sensors in the context of face biometrics. In this section, we evaluate the identification/verification differences of 2.5-D/3-D faces captured by the Kinect and Minolta, which can serve as the reference for the deployment of practical face recognition systems using the Kinect by the state-of-the-art 2.5-D/3-D face recognition algorithms (whose results were reported based on high-quality face scans, such as the data in the FRGC database).

Following the same protocol as described in Section IV-C, we tested different 2.5-D/3-D face recognition algorithms (PCA, LBP, and LGBP-based methods using 2.5-D depth images; ICP and TPS-based methods using 3-D point cloud) on both the KinectFaceDB and the FRGC. For KinectFaceDB, we use all neutral faces in session 1 as the gallery faces and the corresponding neutral faces from session 2 as the probe faces. Similarly, we select two neutral faces from two different sessions of 198 subjects (from FRGC ver.1 [24]) to form the gallery and probe set of FRGC, respectively. In both databases, rank-1 identification rate and verification rate (where $FAR = 0.001$) are reported for comparison.

Table X shows the comparative results of the KinectFaceDB and the FRGC. In the table, the recognition results of FRGC are higher than the results of KinectFaceDB for all tested methods. It should be noted that the result differences in the verification mode are much larger than the ones in the identification mode. This suggests that the Kinect is more appropriate for the noncooperative face identification, whereas a high-quality laser scanner is more suitable for the verification mode which demands more user cooperation. For FRGC, the TPS-based method using 3-D yields better recognition rates than the 2.5-D based methods (i.e., PCA, LBP, and LGBP). On the other hand, for KinectFaceDB, 2.5-D LBP achieves much better recognition rates than the 3-D based methods (i.e., ICP and TPS). This is because the low data quality of the Kinect can significantly deteriorate the sophisticated face registration procedures in ICP and TPS-based methods, and thus greatly deteriorate the final recognition results. This phenomena suggests that simple yet efficient depth descriptors using 2.5-D depth images are preferred for the Kinect-based face recognition in comparison to the methods using sophisticated surface registration based on 3-D points.

Although the 2.5-D/3-D face recognition capability of the Kinect is inferior than the ones of a high-quality laser scanner, its intrinsic advantages make it as a competitive sensor for real-world applications. We summarize its merits for face recognition as follows: 1) it works in real time, which allows online face enrollment in noncooperative scenarios; 2) its 3-D data provides complementary information to the 2-D data, and can be easily integrated for multimodal face recognition; and 3) with streaming 3-D video of the Kinect, 3-D face recognition in video is feasible and can potentially improve recognition rates in comparison to 3-D face recognition based on still images (we have demonstrated the preliminary results in [30]).

Besides its advantages in comparison to the traditional 3-D sensors, it should be noted that the Kinect sensor is limited in dealing with long-distance capturing such as in the surveillance environment. In fact, it is more suitable for face identification/verification tasks in the office/indoor environment for PC operators, or game players. In such circumstances, users are not required to cooperate in the data acquisition procedure. On the contrary, traditional 3-D laser scanners (which is also limited in dealing with long-distance acquisition) require the user to cooperate in the capturing procedure for several seconds in the indoor environment.

With the increasing amount of attention for the Kinect in recent years, we suggest that newly developed 2.5-D/3-D or 2-D + 3-D face recognition algorithms should not only be tested on databases with high-quality scans (such as the FRGC database), but also on the more challenging KinectFaceDB. Robust and reliable face recognition for the Kinect can then be integrated into different applications in more practical scenarios.

VI. Conclusion

In this paper, we presented a complete multimodal (including well-aligned 2-D, 2.5-D, 3-D, and video-based face data) face database based on the Kinect sensor. The database structure and acquisition environment are carefully described. The method of how to obtain the well aligned and processed 2-D, 2.5-D, 3-D, and video face data are thoroughly introduced. We highlighted the advantages of the proposed KinectFaceDB (as well as the Kinect-based face recognition) via the review of existing 3-D face databases and extensive experimental evaluations. In addition, potential applications of the proposed KinectFaceDB are also discussed. Standard face recognition techniques (including PCA, LBP, SIFT, LGBP, ICP, and TPS) are applied on different data modalities (including 2-D, 2.5-D, and 3-D-based face data), and score-level fusion is conducted to demonstrate the performance gain from the integration of depth and RGB. Quantitative comparison (in the context of biometrics) of the proposed KinectFaceDB and the state-of-the-art FRGC database is provided, which can guide the deployment of existing algorithms and the development of new face recognition methods toward more practical systems.

To conclude, the proposed KinectFaceDB supplies a standard medium to fill the gap between traditional face recognition and the emerging Kinect technology. As a future work, it is necessary to revisit the literature on 3-D and 2-D + 3-D face recognition algorithms (which were mostly elaborated with high-quality 3-D face data) using the proposed KinectFaceDB for achieving reliable, robust, and more practical face recognition system using the Kinect. The design of new algorithms and new facial descriptors for the low-quality 3-D data is another important topic to investigate. In addition, 3-D face recognition using video data is a new prospect, where more evaluations will be conducted using the video data from the Kinect in the future. Finally, how to efficiently combine (e.g., via fusion, co-training) different data modalities (RGB, depth, and 3-D) so as to maximize the exploitation of the Kinect for face recognition should also be studied.

Acknowledgment

The authors would like to thank the volunteers from EURECOM for building the database. They would especially like to thank J. Conti, T. Huynh, and Y. Chen for their efforts in constructing the database.

References

[1] (2013, Mar. 31). *Microsoft Kinect* [Online]. Available: http://www.xbox.com/en-US/KINECT

[2] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[3] L. Shao, J. Han, D. Xu, and J. Shotton, Eds., "Special issue on computer vision for RGB-D sensors: Kinect and its applications," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 1295–1296, Aug. 2012.

[4] S. Izadi *et al.*, "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol.*, 2011, pp. 559–568.

[5] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," *Int. J. Robot. Res.*, vol. 31, no. 5, pp. 647–663, 2012.

[6] R. Johnson, K. O'Hara, A. Sellen, C. Cousins, and A. Criminisi, "Exploring the potential for touchless interaction in image-guided interventional radiology," in *Proc. SIGCHI Conf. Human Factors Comput. Syst., CHI'11*, Vancouver, BC, Canada, pp. 3323–3332.

[7] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[8] M. Fallon, H. Johannsson, and J. Leonard, "Efficient scene simulation for robust Monte Carlo localization using an RGB-D camera," in *Proc. 2012 IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 1663–1670.

[9] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. 2011 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Washington, DC, USA, pp. 1297–1304.

[10] M. Gabel, E. Renshaw, A. Schuster, and R. Gilad-Bachrach, "Full body gait analysis with Kinect," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2012.

[11] E. Stone and M. Skubic, "Evaluation of an inexpensive depth camera for in-home gait assessment," *J. Ambient Intell. Smart Environ.*, vol. 3, no. 4, pp. 349–361, Dec. 2011.

[12] J. Preis, M. Kessel, M. Werner, and C. Linnhoff-Popien, "Gait recognition with Kinect," in *Proc. 1st Workshop Kinect Pervasive Comput.*, 2012.

[13] I. Barbosa, M. Cristani, A. Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *Computer Vision—ECCV 2012. Workshops and Demonstrations*, vol. 7583. Berlin, Germany: Springer, pp. 433–442.

[14] B. C. Munsell, A. Temlyakov, C. Qu, and S. Wang, "Person identification using full-body motion and anthropometric biometrics from Kinect videos," in *Proc. 12th Int. Conf. Comput. Vis. ECCV'12*, vol. 7585. Florence, Italy, pp. 91–100.

[15] C. Velardo and J.-L. Dugelay, "Real time extraction of body soft biometric from 3D videos," in *Proc. 19th ACM Int. Conf. Multimedia MM'11*, New York, NY, USA, pp. 781–782.

[16] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, Dec. 2003.

[17] A. Jain, L. Hong, and R. Bolle, "On-line fingerprint verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 302–314, Apr. 1997.

[18] J. Daugman, "How iris recognition works," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 21–30, Jan. 2004.

[19] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[20] P. N. Belhumeur, J. A. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[21] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[22] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[23] J. R. Beveridge *et al.*, "FRVT 2006: Quo vadis face quality," *Image Vis. Comput.*, vol. 28, no. 5, pp. 732–743, May 2010.

[24] P. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. San Diego, CA, USA, Jun. 2005, pp. 947–954.

[25] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2D and 3D face recognition: A survey," *Pattern Recogn. Lett.*, vol. 28, no. 14, pp. 1885–1906, Oct. 2007.

[26] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Comput. Vis. Image Underst.*, vol. 101, no. 1, pp. 1–15, Jan. 2006.

[27] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[28] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image Vis. Comput.*, vol. 10, no. 3, pp. 145–155, Apr. 1992.

[29] (2013, Mar. 31). *KONICA Minolta* [Online]. Available: http://www.konicaminolta.com/

[30] R. Min, J. Choi, G. Medioni, and J. Dugelay, "Real-time 3D face identification from a depth camera," in *Proc. 2012 21st Int. Conf. Pattern Recogn. (ICPR)*, Tsukuba, Japan, pp. 1739–1742.

[31] D. C. Herrera, J. Kannala, and J. Heikkilä, "Joint depth and color camera calibration with distortion correction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2058–2064, Oct. 2012.

[32] D. C. Herrera, J. Kannala, and J. Heikkilä, "Accurate and practical calibration of a depth and color camera pair," in *Computer Analysis of Images and Patterns*, vol. 6855. Berlin, Germany: Springer, 2011, pp. 437–445.

[33] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[34] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1. 2005, pp. 786–791.

[35] G. Medioni and R. Waupotitsch, "Face modeling and recognition in 3-D," in *Proc. IEEE Int. Workshop Anal. Model. Faces Gestures (AMFG)*, 2003, pp. 232–233.

[36] N. Erdogmus and J.-L. Dugelay, "On discriminative properties of TPS warping parameters for 3D face recognition," in *Proc. IEEE/IAPR Int. Conf. Inf. Electron. Vis. (ICIEV)*, Dhaka, Bangladesh, May 2012.

[37] (2013, Mar. 31). *Face Recognition Homepage* [Online]. Available: http://www.face-rec.org/databases/

[38] R. Gross, "Face databases," in *Handbook of Face Recognition*, A. S. Li, Ed. New York, NY, USA: Springer, Feb. 2005.

[39] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[40] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2003.

[41] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Comput. Vis. Image Underst.*, vol. 99, no. 3, pp. 303–331, Sep. 2005.

[42] A. Martinez and R. Benavente, "The AR face database," Dept. Electr. Comput. Eng., Ohio State Univ., CVC, Barcelona, Spain, Tech. Rep. 24, 1998.

[43] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Amsterdam, The Netherlands, Sep. 2008, pp. 1–8.

[44] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis. (WACV)*, Dec. 1994, pp. 138–142.

[45] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Comput. Sci. Dept., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[46] S. Marcel *et al.*, "On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation," in *Recognizing Patterns in Signals, Speech, Images and Videos*. Lecture Notes in Computer Science, vol. 6388. D. Ünay, Z. Çataltepe, and S. Aksoy, Eds. Berlin, Germany: Springer, 2010, pp. 210–225.

[47] T. Faltemier, K. Bowyer, and P. Flynn, "Using a multi-instance enrollment representation to improve 3D face recognition," in *Proc. 1st IEEE Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, Crystal City, VA, USA, Sep. 2007, pp. 1–6.

[48] A. B. Moreno and A. Sánchez, "GavabDB: A 3D Face Database," in *Proc. Workshop Biometrics Internet*, Vigo, Spain, Mar. 2004, pp. 77–85.

[49] "The BJUT-3D large-scale Chinese face database," Multimedia Intell. Softw. Technol. Beijing Municipal Key Lab., Beijing Univ. Technol., Beijing, China, Tech. Rep. MISKL-TR-05-FMFR-001, Aug. 2005.

[50] A. Colombo, C. Cusano, and R. Schettini, "UMB-DB: A database of partially occluded 3D faces," in *Proc. 2011 IEEE Int. Conf. Comput. Vis. Workshops (ICCV)*, pp. 2113–2119.

[51] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authent.*, 1999, pp. 72–77.

[52] S. Gupta, K. R. Castleman, M. K. Markey, and A. C. Bovik, "Texas 3D face recognition database," in *Proc. IEEE Southwest Symp. Image Analysis Interpretation (SSIAI)*, May 2010, pp. 97–100.

[53] (2013, Mar. 31). *3D_RMA: 3D Database* [Online]. Available: http://www.sic.rma.ac.be/∼beumier/DB/3d_rma.html

[54] (2013, Mar. 31). *University of York 3D Face Database* [Online]. Available: http://www-users.cs.york.ac.uk/∼nep/research/3Dface/tomh/3DFaceDatabase.html

[55] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Southampton, U.K., 2006, pp. 211–216.

[56] (2013, Mar. 31). *3dMD* [Online]. Available: http://www.3dmd.com/

[57] (2013, Mar. 31). *Cyberware* [Online]. Available: http://www.cyberware.com/

[58] G. P. Otto and T. K. W. Chau, "'Region-growing' algorithm for matching of terrain images," *Image Vision Comput.*, vol. 7, no. 2, pp. 83–94, May 1989.

[59] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition," *Comput. Vis. Image Underst.*, vol. 101, no. 1, pp. 1–15, Jan. 2006.

[60] I. Kakadiaris *et al.*, "Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 640–649, Apr. 2007.

[61] H. Drira, B. B. Amor, A. Srivastava, M. Daoudi, and R. Slama, "3D face recognition under expressions, occlusions, and pose variations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2270–2283, Sep. 2013.

[62] A. Savran *et al.*, "Bosphorus database for 3D face analysis," in *Biometrics and Identity Management*. Berlin, Germany: Springer, 2008, pp. 47–56.

[63] A. Colombo, C. Cusano, and R. Schettini, "UMB-DB: A database of partially occluded 3D faces," in *Proc. 2011 IEEE Int. Conf. Comput. Vis. Workshops (ICCV)*, Barcelona, Spain, pp. 2113–2119.

[64] J. Duchon, "Splines minimizing rotation-invariant semi-norms in Sobolev spaces," in *Constructive Theory of Functions of Several Variables*, vol. 571. Berlin, Germany: Springer, 1977, pp. 85–100.

[65] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. 2003 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (CVPR)*, pp. 313–320.

[66] F. Matta and J.-L. Dugelay, "Person recognition using facial video information: A state of the art," *J. Vis. Lang. Comput.*, vol. 20, no. 3, pp. 180–187, 2009.

[67] Y.-C. Chen, V. Patel, S. Shekhar, R. Chellappa, and P. Phillips, "Video-based face recognition via joint sparse representation," presented at the *2013 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China.

[68] U. Park, A. Jain, and A. Ross, "Face recognition in video: Adaptive fusion of multiple matchers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[69] M. Hayat, M. Bennamoun, and A. El-Sallam, "Fully automatic face recognition from 3D videos," in *Proc. 2012 21st Int. Conf. Pattern Recognit. (ICPR)*, Tsukuba, Japan, pp. 1415–1418.

[70] M. Hernandez, J. Choi, and G. Medioni, "Laser scan quality 3-D face modeling using a low-cost depth camera," in *Proc. 2012 Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1995–1999.

[71] OpenNI. (2010 Nov.). *OpenNI User Guide* [Online]. Available: http://www.openni.org/documentation

[72] F. Barak, S. Alexander, M. Meir, and A. Yoel, "Depth mapping using projected patterns," U.S. Patent 20100118123, May 2010 [Online]. Available: http://www.faqs.org/patents/app/20100201811

[73] R. Tsai, "An efficient and accurate camera calibration technique for 3D machine vision," in *Proc. Comput. Vision Pattern Recognit.*, 1986.

[74] A. Jain, S. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Biometric Authentication*, vol. 3072. Berlin, Germany: Springer, 2004, pp. 731–738.
[75] A. Dantcheva, C. Velardo, A. D'Angelo, and J.-L. Dugelay, "Bag of soft biometrics for person identification," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 739–777, Jan. 2011.
[76] G. Farinella and J.-L. Dugelay, "Demographic classification: Do gender and ethnicity affect each other?" in *Proc. IEEE/IAPR Int. Conf. Inf. Electron. Vis. (ICIEV)*, Dhaka, Bangladesh, May 2012.
[77] X. Lu, H. Chen, and A. K. Jain, "Multimodal facial gender and ethnicity identification," in *Proc. 2006 Int. Conf. Adv. Biometrics (ICB)*, pp. 554–561.
[78] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision—ECCV 2006*, vol. 3951. Berlin, Germany: Springer, pp. 430–443.
[79] M. Zollhöfer, M. Martinek, G. Greiner, M. Stamminger, and J. Sübmuth, "Automatic reconstruction of personalized avatars from 3D face scans," *Comput. Animat. Virtual Worlds*, vol. 22, nos. 2–3, pp. 195–202, Apr. 2011.
[80] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, Bombay, India, 1998, pp. 839–846.
[81] C. Geng and X. Jiang, "Face recognition using sift features," in *Proc. 2009 16th IEEE Int. Conf. Image Process. (ICIP)*, Cairo, Egypt, pp. 3313–3316.
[82] (2013, Mar. 31). *TABULA RASA* [Online]. Available: http://www.tabularasa-euproject.org/
[83] Y. Huang, Y. Wang, and T. Tan, "Combining statistics of geometrical and correlative features for 3D face recognition," in *Proc. BMVC*, 2006, pp. 90.1–90.10.
[84] 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 *Computer Vision—ACCV 2012 Workshops*, vol. 7728. Berlin, Germany: Springer, 2013, pp. 133–145.
[85] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *Computer Vision—ECCV'98*, vol. 1407. Berlin, Germany: Springer, 1998, pp. 484–498.
[86] I. Matthews, J. Xiao, and S. Baker, "2D vs. 3D deformable face models: Representational power, construction, and real-time fitting," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 93–113, 2007.
[87] J. Harguess and J. Aggarwal, "A case for the average-half-face in 2D and 3D for face recognition," in *Proc. IEEE Comput. Soc. Conf. CVPR Workshops*, Miami, FL, USA, Jun. 2009, pp. 7–12.
[88] Y. Liu, R. Weaver, K. Schmidt, N. Serban, and J. Cohn, "Facial asymmetry: A new biometric," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-01-23, Aug. 2001.
[89] R. Min, A. Hadid, and J. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Proc. 2011 IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops (FG)*, Santa Barbara, CA, USA, pp. 442–447.

Rui Min received the B.Eng. degree in software engineering from Xiamen University, Xiamen, China, in 2008, the M.S. degree in communication and computer security, and the Ph.D. degree in image and signal processing from Télécom ParisTech, Paris, France, in 2010 and 2013, respectively.

He is currently a Postdoctoral Research Associate with the Biomedical Research Imaging Center at the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. He was a Research Assistant at the Multimedia Communications Department of EURECOM from 2010 to 2013. In 2011, he was a Visiting Scholar at the Computer Vision Laboratory of the University of Southern California, Los Angeles, CA, USA. His current research interests include different aspects of face recognition, which contributed to the French national project BIORAFALE and the European project ACTIBIO.

Neslihan Kose received the degree and the M.Sc. degree from the Middle East Technical University, Ankara, Turkey, in 2007 and 2009, respectively, and the Ph.D. degree in image and signal processing from Telecom ParisTech, Paris, France, in 2014.

She is currently a Postdoctoral Research Associate with Telecom ParisTech. She was under the supervision of Prof. J.-L. Dugelay at the Multimedia Department of EURECOM Institute from 2010. During her Ph.D. thesis, she was involved in the European project TABULA RASA.

Jean-Luc Dugelay (M'76–SM'81–F'11) received the Ph.D. degree in information technology from the University of Rennes, Rennes, France, in 1992. His Ph.D. thesis work was undertaken at France Télécom Research (CCETT) at Rennes, France, between 1989 and 1992.

He joined EURECOM in Sophia Antipolis, France, where he is currently a Professor with the Department of Multimedia Communications. His current research interests include domain of multimedia image processing, in particular activities in security (image forensics, biometrics and video surveillance, mini drones), and facial image processing. He has authored or coauthored over 250 publications in journals and conference proceedings, one book on *3-D Object Processing: Compression, Indexing and Watermarking* (Wiley, 2008), five book chapters, and three international patents. His research group was involved in several national and European projects.

Prof. Dugelay has delivered several tutorials on digital watermarking, biometrics, and compression at major international conferences, such as ACM Multimedia and IEEE ICASSP. He has participated in numerous scientific events as a member of the scientific technical committees, Invited Speaker, or Session Chair. He is an Elected Member of the EURASIP Board of Governors. He is/was an Associate Editor of several international journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON MULTIMEDIA, and the Founding Editor-in-Chief of *EURASIP Journal on Image and Video Processing*. He was a member of SPS Image, Video, and Multidimensional Signal Processing Technical Committee from 2002 to 2007, SPS Multimedia Technical Committee from 1999 to 2003 and from 2005 to 2008, and SPS Information Forensics and Security Technical Committee from 2010 to 2012. He coauthored several conference articles that received an IEEE Award in 2011, 2012, and 2013. He co-organized the Workshop on Multimedia Signal Processing (Cannes, 2001), and the 2003 Multimodal User Authentication (Santa Barbara, 2003). In 2015, he will be serving as a General Co-Chair, IEEE ICIP (Qubec City) and EURASIP EUSIPCO (Nice).