3D Face Reconstruction and Multimodal Person Identification from Video Captured Using Smartphone Camera

R. Raghavendra*, Kiran B. Raja*†, Anika Pflug*†, Bian Yang*, Christoph Busch*†
*Norwegian Biometrics Laboratory, Gjøvik University College, 2802 Gjøvik, Norway
†Hochschule Darmstadt - CASED, Haardtring 100, 64295 Darmstadt, Germany
{raghavendra.ramachandra, kiran.raja, anika.pflug2, bian.yang, christoph.busch}@hig.no

***Abstract*—In this paper, we propose a novel approach for reconstructing 3D face in real-life scenarios. Our main objective is to address the most challenging issue that involves reconstructing depth information from a video that is recorded from frontal camera of the smartphone. Such videos recorded using smartphones impose lot of challenges, such as motion blur, non-frontal perspectives and low resolution. This limits the applicability of state-of-the-art algorithms, which are mostly based on landmark detection.**

This situation is addressed with the Scale-Invariant Feature Transformation (SIFT) followed by feature matching to generate consistent tracks. These tracks are further processed to generate a 3D point cloud using Point/Cluster based Multi-view stereo (PMVS/ CMVS). The usage of PMVS/CMVS will however fail to generate a dense 3D cloud points on the weak surfaces of face, such as cheeks, nose and forehead. This issue is addressed by multi-view reconstruction of these weakly supported surfaces using Visual-Hull. The effectiveness of our method is evaluated on a newly collected dataset, which simulates a realistic identification scenario using a smartphone.

***Keywords*—Multiple view reconstruction, 3D Face, Biometrics, Ear recognition, Face recognition**

I. Introduction

Face recognition has received substantial attention from both academia and industry for more than three decades. During the last decade, significant improvements in terms of accuracy were achieved with 2D face images that are captured in controlled scenarios [26]. In practice, however, face recognition still faces a number of challenges, such as pose variation, non-uniform illumination, expressions and occlusions. Recent improvements in the field of face recognition have made an attempt to address some of the above mentioned issues [2], [22].

Especially in surveillance applications, the usage of 3D models for compensating pose variations and non-uniform illumination has gained significant attention. In particular, a number of approaches has been introduced, that reconstruct 3D information from videos. Moreover, 3D face recognition is of paramount interest because of its various applications in the field of computer vision including face recognition, facial expression analysis, avatar, model based image coding and border control applications [2], [9].

However, the success of 3D face reconstruction depends on accurately inferring the depth information from a set of images in order to build a 3D face model. There are wide variety of approaches to build a 3D face model based on 2D images, namely: Active Appearance Model (AAM) [19], [7], [10], Parametrized Appearance Model (PAM) [8] and combined 2D and 3D active appearance model [28].

Even though these approaches appear to be promising, they exhibit drawbacks such as (1) the need for accurate landmark selection, (2) requirement of extensive training, or (3) the need for multiple and properly calibrated cameras. Recent improvements involve creating an accurate 3D face reconstruction from video using either a single image [16], [20] or a single camera [13].

Among these two approaches, single camera based 3D face reconstruction is more appealing, especially for biometrics. The reason for this are (1) it allows a basic form of liveness detection, that makes the system robust against photo attacks. (2) it overcomes the additional computational burden in selecting the best frame from a video to carry out 3D face reconstruction from single image. (3) It also makes additional depth information that is hidden in the spatio-temporal dependence between single video frames explicit, such that it can be processed by the biometric system with the goal of becoming more accurate. (4) With the fact that the video also contains profile views of the person, we can implement a multimodal approach that involves the outer ear as an additional biometric characteristic.

Motivated by these advantages, we consider the problem of 3D face reconstruction using a frontal smartphone camera. More precisely, we use the spatio-temporal dependency of video frames that show a face from multiple perspectives, for computing depth information. The videos were captured by the subjects themselves with the frontal camera of a smartphone. For our test database, we use the Google Nexus S. The resolution of the frontal camera is 0.3 megapixels. For capturing video, the subject holds the smartphone before his face and starts the process of capturing. Then he turns his head from left to the right and back. Hence, the captured videos contain frontal, half profile an full profile yaw-poses with relatively stable roll and pitch. Using these videos, 3D reconstruction is a challenging problem, because there are hardly any stable feature points that are visible throughout the video. Moreover, no ground truth information or training data is available for calibration. The videos also contain motion blur and varying scales, because the subject's hand slightly shakes during video

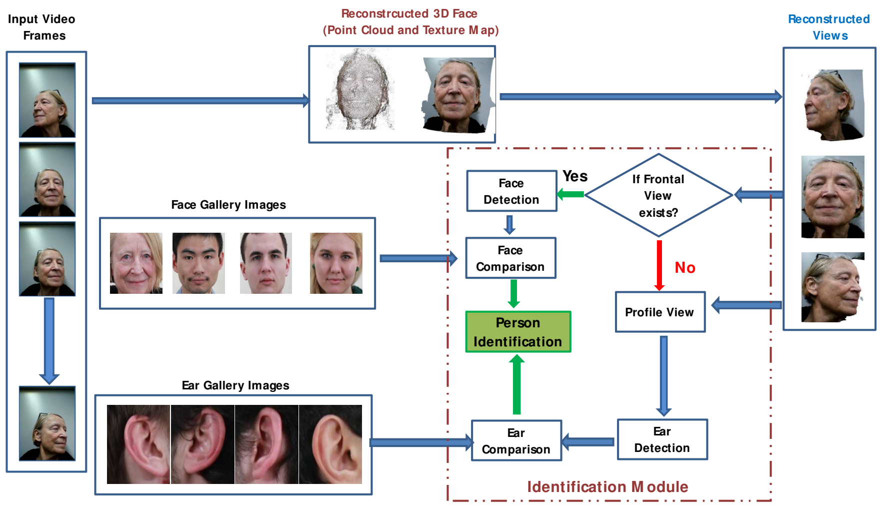

Fig. 1: Overview of the proposed person identification system with 3D face reconstruction from video

capture and the distance between the smartphone and the camera changes throughout the video capturing process.

In this work, we present a system to quickly perform the 3D reconstruction from videos that are captured using the frontal camera of a smartphone. The proposed pipeline is derived largely from existing state of the art schemes. It employs SIFT feature extraction [26], Structure from Motion (SFM) algorithms [14] and Point/ Cluster based Multi-view stereo (PMVS/CMVS) [12]. Using these techniques, we obtain a dense cloud reconstruction of the 3D face.

Due to motion blur, high reflectivity and lack of sufficiently structured texture information, the dense reconstruction of the face from video results in weakly supported surfaces. We propose to address this issue by employing the multi-view reconstruction based on the visual-Hull [15] to accurately reconstruct the 3D face surface. To this extent, we employed the multi-view reconstruction algorithm proposed in [15] that can take the 3D cloud points from PMVS/CMVS and render the strong surface with rich texture information.

The proposed pipeline for this precise application has remarkable advantages, such as: (1) no need of landmark detection and tracking, (2) no need of camera calibration, (3) no need of training data (4) user convenience, because data will be acquired by the user using frontal camera of smartphone (Nexus S). The proposed reconstruction pipeline forms a new concept for multimodal identification on smartphone devices with high user convenience. In this work, we present the results of our first study on such a system, and also explore confronted shortcomings when using existing algorithms.

Thus the main contributions of our work can be summarized as follows: (1) Coupling the available techniques to form a new pipeline to address the application of 3D face reconstruction using a frontal camera of the smartphone (Nexus S). Further, we provide insights on the lessons learnt while building the whole pipeline for this precise application. (2) Extensive evaluation of the proposed pipeline on a newly collected database under realistic conditions and its use in 3D face reconstruction and person identification based on either face or ear. (3) We provide an outlook on possible future applications and research involving our reconstructed 3D models and the collected dataset. (4) To the best of our knowledge, this is the first work that addresses the new and exciting application area of 3D biometric authentication using a smartphone.

The rest of the paper is organized as follows: Section II describes the proposed pipeline for 3D face reconstruction, Section III describes the experimental procedure and results and Section IV draws conclusion and proposals for future research directions.

II. Proposed Scheme

Figure 1 shows the block diagram of the proposed scheme (or methodology) for person identification based on 3D face reconstruction. Given the video, the proposed method begins with feature extraction and matching to accurately construct the tracks. In the next step, these tracks are processed using incremental SFM to compute the 3D cloud points. Since the constructed 3D point clouds fail to accurately render face surface, we propose to use the multi-view reconstruction algorithm based on Visual Hull [15] to accurately construct the 3D face surface and texture. We proceed further to perform the person identification using either face or ear biometric. To perform this, we first check whether we can detect frontal face from the 3D reconstructed face. If yes, we carry out the person identification based on the face, else we perform the person identification using ear that can be detected from the reconstructed 3D face profile. In the following section, we discuss each step involved in building our proposed scheme.

A. Feature extraction and matching

The commonly used feature extraction scheme for the 3D face reconstruction involves locating facial landmark points that outline the structure of eyes, nose, mouth, eyebrows and the facial boundary. However, accurate landmark detection and tracking requires [21] extensive training and robustness to various viewpoints (both full profile and half profile). These drawbacks will limit the applicability of the 3D face reconstruction schemes (like AAM, PAM, etc.) based on landmark point detection and tracking especially for our precise application of 3D face reconstruction from video captured using smartphone. Hence, we employed the Scale-Invariant Feature Transformation (SIFT) [26] to extract the features from each frame of the recorded video. In this work, we employed the GPU implementation of SIFT by considering its speed and user friendly interface [26]. Here, the choice of SIFT feature extraction looks appealing for our precise application by considering its robustness against: (1) image resolution (as we are processing low resolution video with 0.3 megapixel) (2) various viewpoints (3) self-occlusion and (4) clutters.

In the next step, we carry out the feature matching between images based on the nearest neighbor search [6]. The obtained matches are then pruned and verified using RANSAC based estimation of the fundamental or essential matrix [23] [4]. Finally, we combine all matches into tracks to form a connected set of key points across multiple images. In order to prune out the inconsistent tracks, we perform the track selection that contains at least two key points for the next stage that involves in 3D cloud reconstruction.

B. 3D surface reconstruction

After obtaining the consistent tracks, we proceed further to run the structure from motion algorithm for recovering a 3D location for each track. This can be carried out by using bundle adjustment, which minimizes the re-projection error between each tracks and its corresponding image [4]. In order to overcome the problem of SFM getting struck in a local minima [23], we carry out an incremental reconstruction procedure by adding one frame at a time. In this work, we employed the VisualSFM [27] an open source GUI that integrates three packages namely: SIFT on GPU (siftGPU), incremental SFM system and multi core bundle adjustment [27]. Further, VisualSFM runs very fast by exploiting the multi-core acceleration for feature extraction, matching and bundle adjustment. In addition, Visual SFM also integrates an interface to run both PMVS/CMVS for 3D dense reconstruction. Hence, given the video sequence, we run the VisualSFM to get the 3D dense reconstruction of the face surface.

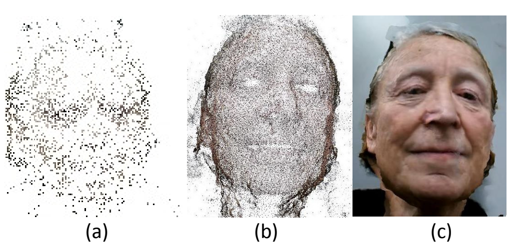

Fig. 2: Illustration of 3D Clouds obtained using (a) PMVS/CMVS (b) Visual-Hull (c) after texture rendering

Figure 2(a) shows the 3D cloud point obtained using PMVS/CMVS using VisualSFM. Here, it can be observed that, the use of PMVS/ CMVS will fail to construct the strong surfaces for the 3D face reconstruction. Thus, in order to overcome this problem, we employed the multi-view reconstruction algorithm based on Visual-Hull [15] to reconstruct the difficult face surfaces (specially cheeks, nose and ear) that are not sampled densely using PMVS/CMVS. The main idea of the visual hull algorithm involves in computing the free space support that detects the highly supported surface boundary and reconstruct the weak surface by optimizing the t-weights in an s-t graph [15]. Thus the employed visual hull algorithm [15] performs the multi view reconstruction by accepting the 3D cloud surfaces generated from PMVS/CMVS. Figure 2(b) shows the multi-view reconstruction of the weak face surfaces especially, cheeks and nose region while Figure 2(c) shows the final reconstructed image after texture rendering based on the point clouds [15].

C. Face/Ear based person identification

After accurately reconstructing the 3D head model from a video, we proceed further to identify the subject by comparing the frontal view of the reconstructed face with the enrolled samples. We are also addressing the 3D profile (half and/or full) face reconstruction. Hence, we are also interested to explore the contribution of ear recognition to the system's overall robustness. To this extent, we compare the outer ears, in cases where the accurate 3D frontal face reconstruction failed. In this way, we are making an attempt to explore the multi-modality from the reconstructed 3D structure to identify a person (or subject) based on either face or ear.

We collected the enrollment samples in a controlled illumination (or studio) environment using Canon EOS 550D DSLR camera. The usage of two different cameras and the fact that the enrollment images and the probe images were taken under different conditions, simulating real-life intra/inter-class variance between the images. For each subject, we captured one frontal view for face with a neutral expression and two full profile views that represent left and right ear. We then perform a series of pre-processing steps that include detecting and cropping a face region using Viola-Jones face detector [24].

Fig. 3: Examples of the enrolled samples (a) Enrolled samples (b) Corresponding face images

The detected face region is then resized to 120×120 pixels to reduce the computation cost. Finally, we apply a Gaussian filter ($\sigma = 2$) to remove noise high frequency noise. A similar procedure is followed for extracting the the ear region from the enrollment images. The ear detector was trained using OpenCV 2.4 with 2000 positive and 300 negative samples. The positive training images were taken from manually cropped

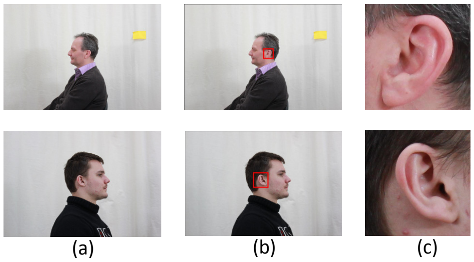

Fig. 4: Examples of enrolled Ear samples (a) Enrolled samples (b) Ear detection results (c) Corresponding Ear images

images from the publicly available UND-J2 database [3], the IIT-Kanpur database [18] and the AMI ear database [1]. The negative training samples were chosen randomly from the INRIA person detection image set [11]. Figure 3 shows the example of the enrolled samples and their corresponding processed face samples while Figure 4 shows the enrolled and processed ear samples. In order to perform the person identification, we compare the reconstructed 3D face with the enrolled sample of the subject, we first correct the pose of all reconstructed faces towards a frontal view to have yaw, pitch and roll values to zero. Then we project the head model to the 2D plane. In this work, we compare the frontal view from the reconstructed 3D face (after projecting to 2D) with the gallery sample (which is also a frontal view) using Sparse Representation Classifier (SRC) [25].

Fig. 5: Illustration of multi-view reconstruction using proposed scheme

A similar procedure is also followed to compare the ear detected from reconstructed 3D profile face with the enrolled sample of the ear. Hence, in this work, we compare either left or right (not both) ear with the corresponding left or right enrolled sample using Sparse Representation Classifier (SRC) [25]. In this work, we choose the SRC by considering its robustness and performance on both face [25] and ear [17] biometrics. Other existing classifiers such as SVM, Kernel Discriminant analysis and etc. can also be used at this stage. However, the study on these classifiers remains beyond the scope of this work which focuses mainly on the 3D face reconstruction using smartphone camera. The ear to compare is the first ear that is visible in the video stream (see next section for details).

III. Experimental Results

In this section, we describe both qualitative and quantitative results of our proposed pipeline for fast and accurate 3D face reconstruction and person identification using the frontal camera of Google Nexus S smartphone. First we discuss our new face video dataset. We then present the qualitative results of the proposed 3D face reconstruction scheme and finally, we presents the quantitative results of the multimodal person identification, which includes face and ear recognition. Our

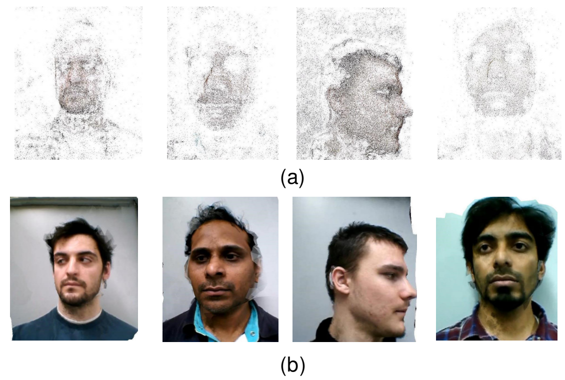

Fig. 7: Illustration of (a) 3D Cloud points (2) Corresponding 3D image rendered using proposed scheme

newly collected dataset consists of 25 subjects among which there are 19 males and 6 females. Each subject is asked to rotate his head by holding the smartphone in hand. This particular setting for the data collection appears quite appealing from the user's point of view, however it poses large number of challenges for the accurate 3D face recognition. The recorded video from the frontal camera of Google Nexus S smartphone is of moderate quality with 0.3 megapixel VGA, which makes it further interesting and challenging. Every subject is asked to record a video of himself, where he rotates his head from frontal pose to the left, the right and back to the frontal pose. The subjects could chose by themselves, whether they turn their heads left or right first. Each video recording takes between 2 -3 seconds. We then decompose each of these videos into its corresponding frames before performing the 3D face reconstruction. Thus, the collected dataset has any one of the following rotation patterns for each subject:(1) rotation of head starting from frontal face till left full profile face (2) rotation of head starting from frontal face till right profile face (3) rotation of head starting from left full profile till right full profile or vice versa.

Figure 6 shows the 3D reconstruction of the half profile face. As it can be clearly seen, our method manages to reconstruct the shape of the cheeks and the outer ear shape. Moreover, we see that, one can simulate self-occlusion by adjusting the reconstructed model to the according pose.

Figure 7 illustrates the qualitative results of the proposed pipeline for 3D reconstruction that shows both 3D point cloud and corresponding 3D frontal face after texture rendering. Figure 5 illustrates the full profile reconstruction of the rotated face. These results further justify the efficiency of the proposed pipeline for 3D reconstruction. The proposed scheme has

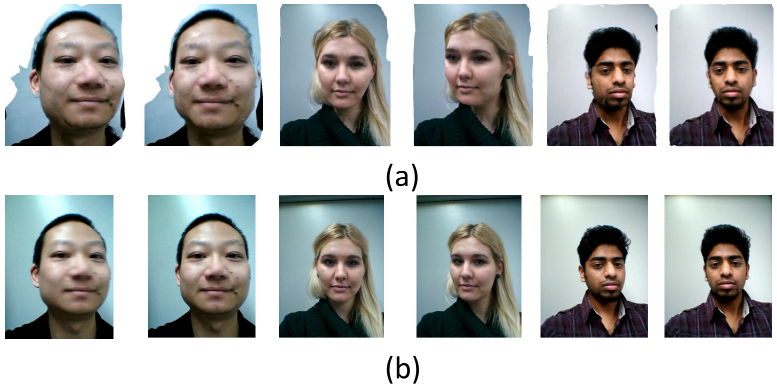

Fig. 6: Results for 3D reconstruction (a) 3D reconstruction (b) Corresponding video frame

shown a remarkable performance in terms of reconstructing 3D frontal view on 17 subjects out of 25 subjects in total. For the remaining 8 subjects, the proposed scheme can accurately reconstruct the 3D profile view. One possible reason is the fact that the profile can be reconstructed, even though the reconstruction of the frontal view failed due to lack of frontal face images. This can happen, if the subjects have moved their head too fast.

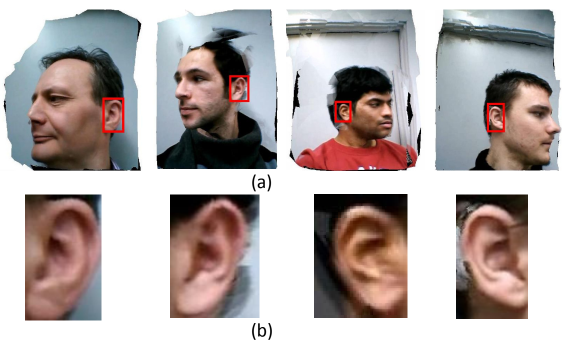

Fig. 8: Illustration Ear detection from reconstructed 3D profile face image (a) Ear detection results (b) Corresponding segmented ear image (2D)

Figure 8 shows the ear detection and localization from the reconstructed 3D profile view. The person identification is carried out by comparing the projected ear (to 2D) with its corresponding gallery samples. In order to provide the comprehensive comparison, we manually choose the profile sample from the video frame and compare the same with the gallery samples.

Table I shows the performance of the face recognition on using frontal view from the 3D face and frontal view of the face from a recorded video. In this work, all the results are presented in terms of closed-set identification rate (rank 1) that is obtained by comparing $1 : N$ subjects in the dataset, therefore, a higher value of identification rate corresponds to better accuracy of the algorithm. For each subject, we acquired three enrolment samples under controlled illumination conditions, namely: frontal face with neutral expression, full profile with left ear and full profile with right ear. Thus, for the comparison, we first check for the existence of a frontal face and, if we are able to detect it, we compare the reconstructed frontal view to the frontal view from the gallery. Otherwise, we compare the detected and extracted ear from the reconstructed profile view with its corresponding ear from the enrolment samples (either left or right).

We have presented the results of the proposed scheme for both unimodal and multimodal recognition using face and ear. For the comprehensive comparison, we have also compared the quantitative results of the proposed 3D reconstruction scheme with the 2D video frame. It can be observed from the Table I that, the proposed scheme shows the best performance among the possible combinations, with an identification rate of 82.25% on the frontal face and 75% on the ear. Furthermore, the proposed method also shows an outstanding performance of 80% when using both face and ear images, by indicating an improvement of 12% as compared to the 2D video frame. The degraded performance of the 2D video can be attributed to the presence of noises due to motion blur, illumination etc. The multimodal results reported in this paper are obtained by associating the correctly identified subjects from both frontal face and ear modality (OR rule). These quantitative results further justify the applicability and efficiency of the proposed scheme. In this work, we carried out the 2D comparison because of the

Modality	Methods	No. of subjects	Identification Rate (%)
Unimodal (Face / Ear)	3D face	17	82.35
	Video face	17	70.58
	3D ear	8	75.00
	Video ear	8	62.50
Multimodal (Face and Ear)	3D face + 3D ear	25	**80.00**
	Video face + Video ear	25	**68.00**

TABLE I: Qualitative performance of the proposed scheme

lack of 3D enrollment data. Even though one can argue that, the use of 2D video can also provide equally good results, but, it is well know that, the 2D based biometric systems are highly vulnerable for attacks (eg. Photo attacks) and hence fails to prevents spoofing. While the use of 3D is very difficult to spoof as compared to that of 2D [5]. Thus, the proposed scheme opens up more avenues for secured applications using smartphones. Kindly refer http://youtu.be/VQTGh5AjM38 for additional comprehensive results.

IV. Conclusion

We have presented a new scheme for 3D reconstruction and recognition from videos that are acquired using the frontal camera of the Google Nexus S smartphone. This scheme can be seen as a proof of concept for future authentication systems, which are more accurate and more robust against spoofing than existing approaches. The proposed scheme is derived from the available techniques, which are coupled in a novel way to achieve a significantly better performance in 3D reconstruction and accurate person recognition.

The proposed scheme stands different when compared to the state of art schemes by overcoming the requirement of landmark points detection, exhaustive training, camera calibration and multiple cameras. Further, the proposed scheme is extensively evaluated on a newly collected database comprising of 25 subjects by considering real-life scenarios. The proposed scheme has shown good performance in reconstructing 3D frontal faces from 17 different subjects out of 25.

Our further work involves improving the system to achieve better reconstruction and matching speed, selection of key frames for improving the overall speed and also on exploring the surface reconstruction scheme to build more robust and accurate system. Moreover, we expect further performance improvement by including both ears into the pipeline.

Acknowledgment

This work was funded by the EU 7^{th} Framework Program (FP7) under grant agreement n^o 284862 for the large-scale integrated project FIDELITY.

References

[1] Ami ear database:-. "http://www.ctim.es/research_works/ami_ear_database/#cita".

[2] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2d and 3d face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007.

[3] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon. A survey on ear biometrics. *ACM Computing Surveys (CSUR)*, 45(2):22, 2013.

[4] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 72–79. IEEE, 2009.

[5] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–7. IEEE, 2011.

[6] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923, 1998.

[7] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10):1380–1384, 2004.

[8] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[9] C. Busch and M. Brauckmann. Towards a more secure border control with 3d face recognition. *Norsk informasjonssikkerhetskonferanse (NISK)*, 2012, 2012.

[10] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor. Coupled-view active appearance models. In *Proceedings of the British machine vision conference*, volume 1, pages 52–61, 2000.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, 2005.

[12] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(8):1362–1376, 2010.

[13] T. Hara, H. Kubo, A. Maejima, and S. Morishima. Fast-accurate 3d face model generation using a single video camera. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1269–1272. IEEE, 2012.

[14] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[15] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3121–3128. IEEE, 2011.

[16] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):394–405, 2011.

[17] A. Kumar and T.-S. T. Chan. Robust ear identification using sparse representation of local texture descriptors. *Pattern Recognition*, 2012.

[18] A. Kumar and C. Wu. Automated human identification using ear imaging. *Pattern Recogn.*, 45(3):956–968, 2012.

[19] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[20] M. Pamplona Segundo, L. Silva, and O. R. P. Bellon. Improving 3d face reconstruction from a single image using half-frontal face poses. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1797–1800. IEEE, 2012.

[21] U. Park and A. Jain. 3d model-based face recognition in video. *Advances in Biometrics*, pages 1085–1094, 2007.

[22] A. Scheenstra, A. Ruifrok, and R. Veltkamp. A survey of 3d face recognition methods. In *Audio-and Video-Based Biometric Person Authentication*, pages 325–345. Springer, 2005.

[23] P. Torr and A. Zisserman. Robust computation and parametrization of multiple view relations. In *Computer Vision, 1998. Sixth International Conference on*, pages 727–732. IEEE, 1998.

[24] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137 – 154, May 2004.

[25] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[26] C. Wu. SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). "http://cs.unc.edu/~ccwu/siftgpu", 2007.

[27] C. Wu. Visualsfm: A visual structure from motion system. 2011.

[28] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+3d active appearance models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2. IEEE Computer Society; 1999, 2004.