FACE-BASED ACTIVE AUTHENTICATION ON MOBILE DEVICES

Mohammed E. Fathy, Vishal M. Patel, Rama Chellappa

Center for Automation Research, University of Maryland, College Park, MD 20742
{mefathy, pvishalm, rama} *(at)* umiacs.umd.edu

ABSTRACT

As mobile devices are becoming more ubiquitous, it becomes important to continuously verify the identity of the user during all interactions rather than just at login time. This paper investigates the effectiveness of methods for fully-automatic face recognition in solving the Active Authentication (AA) problem for smartphones. We report the results of face authentication using videos recorded by the front camera. The videos were acquired while the users were performing a number of tasks under three different ambient conditions to capture the type of variations caused by the 'mobility' of the devices. An inspection of these videos reveal a combination of favorable and challenging properties unique to smartphone face videos. In addition to the variations caused by the mobility of the device, other challenges in the dataset include partial faces, occasional pose changes, blur and face/fiducial points localization errors. We evaluate still image and image set-based authentication algorithms using intensity features extracted around fiducial points. The recognition rates drop dramatically when enrollment and test videos come from different sessions . We will make the dataset and the computed features publicly available to help the design of algorithms that are more robust to variations due to factors mentioned above.

Index Terms— Face recognition, mobile devices, active authentication, biometrics recognition.

1. INTRODUCTION

Developments in sensing and communication technologies have led to an explosion in the use of mobile devices such as smartphones and tablets. Mobile devices make the management of personal information such as emails, bank accounts and profiles convenient and flexible. However, with the increasing use of mobile devises one has to constantly worry about the security and privacy as the loss of a mobile device would compromise personal information of the user.

Most mobile devices use passwords, pin numbers, or secret patterns for authenticating users. As long as the device remains active, there is no mechanism to verify that the user originally authenticated is still the user in control of the device. As a result, unauthorized individuals may improperly gain access to personal information of the user if the password is compromised. Active Authentication (AA) systems deal with this issue by continuously monitoring the user identity after the initial access has been granted. However, AA remains an unsolved problem specially for smartphones. Various efforts for authenticating smartphones have been proposed. Examples include systems based on screen touch gestures [1, 2], gait recognition [3], and device movement patterns (as measured by the accelerometer) [4]. As smartphones come equipped with a user-facing camera and multiple core processors/GPUs, it is becoming more feasible to utilize the existing body of research in face recognition for face-based AA on smartphones.

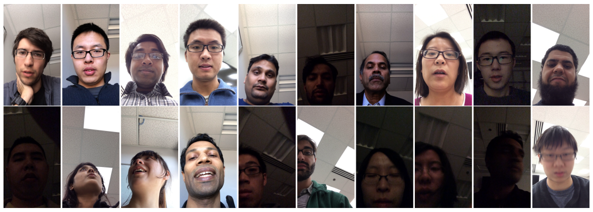


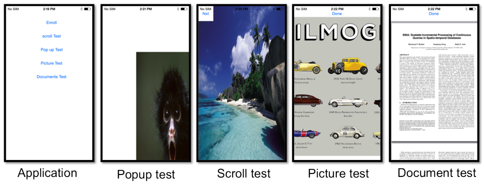
Fig. 1: Sample video frames for 20 (out of 50) users. The head of the user is close always close to the camera. The bottom row shows some of the challenges present in the data including illumination, pose, expression, partial faces and blur.

Over the years, many algorithms have been proposed for face recognition from still-images, image-sets and videos. Examples include Eigenfaces [5], Fisherfaces [6, 7], SRC [8], AHISD/CHISD [9], SANP [10], DFRV [11], and MSSRC [12] just to name a few. While such algorithms have been tested on challenging benchmarks [13, 14, 15, 16] it is hard to predict if they will achieve the same performance on smartphone face videos as they may involve challenges different from those in surveillance-based face recognition datasets. Thus, it becomes necessary to (a) build a dataset that captures the challenges of smartphone face videos and (b) provide a benchmark to quantify how well existing algorithms can solve the problem in addition to helping future research efforts. MOBIO is the only other benchmark that is based on smartphone face videos [17]. Unlike our study, the benchmark of MOBIO considered only still-image-based methods and only one frame per video is manually cropped, normalized and included in the evaluation [18]. So challenges such as partial faces and incorrect facial/fiducial point detections are not addressed in that work.

In this paper, we present a benchmark for measuring (and comparing) the effectiveness of face recognition techniques when used for active authentication using face videos captured by the smartphone's front-facing camera. The benchmark dataset consists of 750 videos from 50 different users and two evaluation protocols that reflect some of the challenges a typical face-based active authentication system is likely to deal with in practical smartphone applications. We used the two protocols to evaluate several existing techniques for still-image-based and image-set-based face recognition including state-of-the-art ones. Although some techniques perform better than others, the best performance obtained is still not adequate even when the features are extracted around face fiducial points. To encourage further research, we will make the dataset, the extracted features and evaluation protocols publicly available.

The paper is organized follows. In Section 2, we describe the dataset collected by the authors' group. The preprocessing pipeline including face detection and feature extraction is explained in Sec-

This work was supported by cooperative agreement FA8750-13-2-0279 from DARPA.



Application Popup test Scroll test Picture test Document test

Fig. 2: Screen shots of the application and tasks used to collect data on an iPhone 5s.

tion 3. Section 4 presents the evaluation protocols while Section 5 gives the benchmark results. The paper is concluded in Section 6 with a brief summary and discussion.

2. MOBILE FACE DATASET DESCRIPTION

The dataset was collected using a custom-written app on an iPhone 5s. The app collected data for five different tasks (See Fig. 2). During each task, the app recorded each users' face video from the front camera as well as the touch data sensed by the screen[1]. Each user performed five tasks in three settings (sessions) with very different environmental conditions. These setting were as follows: (a) in a well-lit room, (b) in the same room but with dim lighting, and (c) in a different room with natural daytime illumination. Although the three sessions of a given user were collected in the same day, the benchmark results indicate that the dataset is still challenging as state-of-the-art methods fail to achieve good performance in cross-session evaluations. The different tasks are described below.

- **Enrollment Task**: The user would enroll his/her face by turning their head to the left, then to the right, then up, and finally down while being recorded by the front-facing camera on the iPhone. Following the enrollment task, the user would perform four tasks with both face and screen touch data being recorded simultaneously. The four tasks are described as follows.
- **Document Task**: The user is presented with a 12-page long PDF research paper and is asked to count the number of items indicated by the test proctor such as figures, tables etc.
- **Picture Task**: A large poster-like image displayed 72 cars with different colors in a 12 by 6 table. The user was asked to count the number of cars of a particular color selected by the test proctor. Only a few cars could be seen at any given time on the screen and so scrolling was necessary to view all cars.
- **Popup Task**: 15 images were positioned off screen in such a way that only a little bit of the image was shown. The user was required to drag the image and position it in the center of the iPhone to the best of their ability.
- **Scrolling Task**: The app displayed a collection of images that were arranged horizontally and vertically. Each image would take up the whole screen and the user was required to swipe (using their finger) on the screen left and right or up and down in order to navigate through the images.

The new dataset consists of 750 video sequences from 50 different users. Before starting each task, the task description was verbally conveyed to the user. No further instructions were given to the users regarding their pose or the way they held and interacted with the device while doing the different tasks. The resolution of each video is 1280 × 720. The average video duration is 11 seconds for the Enrollment Task, 43 seconds for the Document Task, 40 seconds for the Picture Task, 51 seconds for the Popup Task, and 32 seconds for the Scrolling Task. Figure 1 shows some sample recorded images from this dataset.

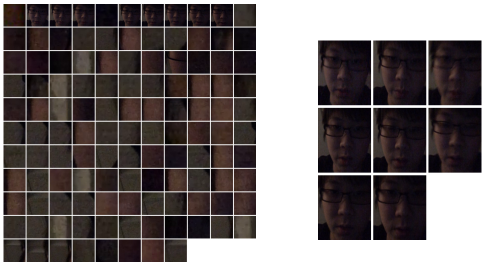


Fig. 3: Increasing the size of the smallest search window of VJ detector to 25% of the frame size eliminates all the false alarms within the 149 detections (shown in the left) made in a sample video file while keeping the 8 true positives (shown in the right).

An inspection of videos in this dataset reveals a combination of characteristics that is unique to front camera videos. Some of these are favorable characteristics that can be utilized to increase the robustness and efficiency of the authentication process. For example, most users keep their heads close to the smartphone while using it. Most of the time, users keep their faces and eyes directed towards the phone (i.e. the camera) while they interact or read something off the phone although they may turn their heads occasionally, for example, to speak to someone or look around.

Other characteristics present in these videos are challenging for many state-of-the-art face authentication systems. The fact that the device (and so the camera) is held by the user during data acquisition phase contributes to many observed variations in face images. For example, the imaging device is subject to shakes and sudden movements which result in blurred frames in some of the videos (even normal head movements contribute to the blurring of faces). Users can also adjust the height and distance of the device relative to their heads in the middle of any interaction, which can change the background and the location, size and distortion of the face within the images. We also noticed that some users hold the device during some interactions such that only a part of the face remains fully within the field of view of the camera.

In addition to the aforementioned challenges, a major challenge with smartphone face videos is the variations in illumination and contextual conditions within the videos of the same subject resulting from the mobility of the device. This issue is practically inevitable as smartphones are designed to be carried and used everywhere and all the time. To capture this mobility challenge in our dataset, the data for each user has been collected under the three aforementioned sessions, each of which has different illumination condition.

3. PREPROCESSING

Face Detection The first step is to locate the user's face from each frame. While there are several algorithms for face detection [19], we used the Viola-Jones (VJ) detector [20] as it is relatively fast and has tuned open-source implementations available on popular smartphone platforms. We utilized the fact that the user's face is close to the camera during acquisition time by setting the size of the smallest search window to 25% of the frame resolution. This makes the detector run 46 times faster (28 fps on MATLAB using a single-core 2.2 GHz processor) while reducing the false positives drastically which

[1] Note that we only focus on the face data in this paper. Results on the touch data will be presented elsewhere.

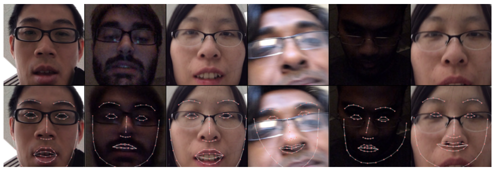


Fig. 4: Top row: cropped facial detections (before histogram normalization). Bottow row: the fiducial points computed by the pretrained model of [21]. The left three pairs are examples of good results while the right three pairs are examples of incorrectly placed fiducial points.

usually have small dimensions (see Figure 3 for an example).

It is worth noting that some frames contribute no detections. In many cases, this is because of partial faces or the user looking away from the phone.

Fiducial Point Detection Given the face bounding box, we use the pre-trained landmark detector of [21] available from [22] to identify fiducial points at the eyes, nose and mouth. We use these to guide the feature extraction step in an effort to normalize appearance variation due to pose and expression. For robustness, we drop any detection if we find that any of the fiducial points on the eyes, nose or mouth is outside or too close to the boundary of the face detection rectangle. A fiducial point is considered too close if it lies less than 5 pixels away from any of the four sides of the detection rectangle. Since all preprocessing is fully automatic, the resulting detections may not always be perfect. Figure 4 shows examples of good and bad results obtained. We do not attempt to filter out these bad results manually and we rely on the robustness of the subsequent image-set classifier to deal with such outliers.

The detected faces are then cropped out and rescaled to 256×256. We then apply histogram equalization to reduce the variations due to illumination. The resulting face images are then used for feature extraction.

Feature Extraction Given a detected face image $\mathbf{I}$, we extract a 400D feature vector $\boldsymbol{x} = \mathbf{F}(\mathbf{I})$. We consider two types of intensity features $\mathbf{F}_1$ and $\mathbf{F}_2$. The first type $\mathbf{F}_1$ is holistic in nature which works by rescaling $\mathbf{I}$ into 20×20 and arranging the intensity values into $\boldsymbol{x}$. The second type $\mathbf{F}_2$ utilizes the locations of fiducial points to improve the alignment of the intensity values in $\boldsymbol{x}$. It achieves this by computing four bounding boxes of the mouth, left eye, right eye, and nose fiducial points and then we extend each bounding box by including 5 more pixels in each direction to include more context. Subsequently, we resize the mouth box to 7×14, each eye box to 9×11, and the nose box to 8×13. This gives a total of 400 intensity values which are arranged into the feature vector x (see Figure 5 for illustration). We refer to such features as *MEEN* features because they are constructed from the Mouth, left Eye, right Eye, and the Nose. As expected, we obtain better accuracy using the MEEN features.

If there are n face images $\{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_n\}$ in a given video $\mathbf{V}$, we obtain a set of n corresponding feature vectors $\mathbf{F}(\mathbf{V}) = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$. Image-set-based face recognition techniques (or simple extensions of still-image-based ones) are then used for training and/or testing.

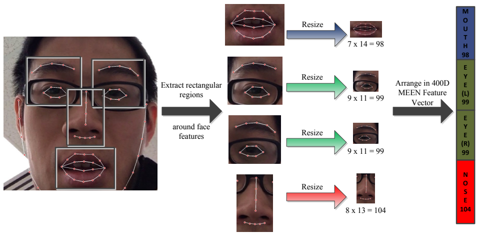


Fig. 5: MEEN features. The regions surrounding the landmarks on the mouth, eyes and nose are extracted, rescaled and arranged into a 400D feature vector.

4. EVALUATION PROTOCOLS

A typical practical scenario for using an active face authentication system on smartphones would involve an enrollment stage in which the user enrolls their face for at least one session. After enrollment, the system is set to query mode and is expected to receive a continuous sequence of image-set queries where the overall amount of query data is much larger than the enrollment data. Since smartphones are designed to be used everywhere, the query sets may involve places and illumination settings different from those present during enrollment.

We consider in this benchmark two evaluation protocols that model this scenario. In both protocols, the overall amount of query data is bigger than that of enrollment data. In addition, the illumination settings are different from those of enrollment. In protocol 1, the system is trained on the enrollment videos from one session (e.g. session 1) and is tested on non-enrollment video clips from the other two sessions (e.g. sessions 2 and 3). In protocol 2, the system is trained on the data from two enrollment sessions (e.g. sessions 1 and 2) and is tested on non-enrollment video clips from the other session (e.g. session 3).

The test video clips are created from the non-enrollment task videos by splitting each task video into 10-second long video clips and keeping only those clips with at least one face detection. The rationale is that in practice, the system should authenticate the user continuously and one way to achieve this is to run a query periodically. The query period we have adopted in this work is 10 seconds. Given a query video clip, the system should identify the subject present in that video clip. Accordingly, we cast the problem as a 50-class identification problem.

5. EXPERIMENTAL RESULTS

We evaluated 4 still-image-based methods including Eigenfaces (EF) [5], Fisherfaces (FF) [6, 7], Large-Margin Nearest Neighbour (LMNN) [23], and Sparse Representation-based Classification (SRC) [8]. In addition, we included 5 image-set-based methods based on Affine Hull-based Image Set Distance (AHISD) [9], Convex Hull-based Image Set Distance (CHISD) [9], Sparse-Approximated Nearest Points (SANP) [10], Dictionary-based Face Recognition from Videos (DFRV) [11], and Mean-Sequence SRC (MSSRC) [12]. We adjusted the computation of the data mean and scatter matrices in EF and FF by reweighting the contribution of the samples of each class so that all classes contribute equally regardless of the different class sizes. As in [6], we dropped the first three principal components in EF and use the subsequent 150 components to define the PCA projection matrix (adding more components does not improve the recognition rate in our experiments). Still-image-based methods process an image-set query by independently classifying each vector in the query and declaring the most frequently occuring label as the winner.

Table 1 shows the recognition rates under protocol 1 and Table 2 shows the recognition rates under protocol 2. For the sake of com-

Table 1: Recognition rates under protocol 1: The different models are trained using one session's enrollment videos and tested on video clips from another session. For each row, we show **in bold** the three highest recognition rates achieved for this experimental setting.

Enrollment Sessions	Testing Sessions	EF	FF	LMNN	SRC	AHISD	CHISD	SANP	DFRV	MSSRC
1	2	40.95	**54.48**	30.80	**52.79**	22.17	14.55	17.26	29.78	**47.21**
1	3	34.02	**45.27**	30.77	**51.18**	21.30	13.91	17.01	35.65	**46.15**
2	1	22.23	25.52	13.41	**44.18**	10.23	7.97	10.60	**32.55**	**43.06**
2	3	49.70	**56.80**	43.05	**58.58**	47.78	44.67	44.97	46.30	**60.36**
3	1	**28.05**	**24.77**	**22.05**	17.64	10.69	11.63	13.04	19.89	17.64
3	2	**55.50**	**56.01**	50.76	**51.95**	46.87	41.12	43.82	47.04	45.85

Table 2: Recognition rates under protocol 2: The different models are trained using the enrollment videos of two sessions and tested on video clips from the remaining session. For each row, we show **in bold** the three highest recognition rates achieved for this experimental setting.

Enrollment Sessions	Testing Sessions	EF	FF	LMNN	SRC	AHISD	CHISD	SANP	DFRV	MSSRC
{1, 2}	3	55.18	**74.85**	48.37	**72.93**	51.18	47.04	48.08	52.81	**72.19**
{2, 3}	1	**30.11**	**54.69**	25.33	24.20	14.35	16.51	16.79	**39.21**	22.14
{1, 3}	2	63.96	**71.91**	56.18	**72.93**	50.08	43.15	46.70	50.25	**69.71**

Table 3: Recognition rates when enrollment videos and non-enrollment test video clips come from the same session. The recognition rates for such setting are relatively good compared to those of protocol 1 and protocol 2. For each row, we show **in bold** the three highest recognition rates achieved for this experimental setting.

Enrollment Sessions	Testing Sessions	EF	FF	LMNN	SRC	AHISD	CHISD	SANP	DFRV	MSSRC
1	1	91.84	93.53	**94.65**	93.25	94.00	**95.50**	**94.84**	91.56	93.25
2	2	79.70	84.77	84.94	84.94	**86.46**	**85.45**	85.11	83.59	**85.45**
3	3	82.25	**86.98**	**83.58**	80.47	**85.06**	82.54	83.14	76.78	73.52
{1, 2}	1	92.31	93.34	**94.18**	92.96	93.90	**95.31**	**94.47**	92.03	93.06
{1, 2}	2	81.73	83.76	83.76	**85.11**	**85.96**	**85.11**	84.77	83.93	**85.79**
{2, 3}	2	81.90	**84.09**	82.57	79.86	**85.45**	**84.43**	83.59	82.57	68.02
{2, 3}	3	84.17	**91.72**	85.21	82.40	**88.46**	84.76	**85.50**	81.66	73.22
{1, 3}	1	93.25	93.25	93.71	92.78	**94.09**	**96.06**	**95.03**	92.50	92.68
{1, 3}	3	83.14	**87.43**	83.88	**85.36**	**84.62**	82.40	83.58	78.25	83.88

pleteness, we show in Table 3 the recognition rates obtained by testing on the non-enrollment video clips from the same sessions used for training. These are the accuracies that would be obtained when the mobility challenge is excluded (although other challenges such as partial faces, blur, expression, pose variations, and face/landmark localization errors are still present). All tables show results obtained using the fiducial point-based features. The less superior results obtained with holistic features are not shown due to page limitations.

Tables 1 and 2 indicate that the best performing methods are FF, SRC, and MSSRC. Yet, the recognition rates (in percentages) they achieve for protocol 1 range between 24.8 and 56.8 for FF, 17.6 and 58.6 for SRC, and 17.6 and 60.4 for MSSRC. For protocol 2, they range between 54.7 and 74.9 for FF, 24.2 and 73.9 for SRC, and 22.1 and 72.2 for MSSRC. Compared to the recognition rates obtained in 3, it can be seen that the evaluated methods (including state-of-the-art image-set methods) have difficulty coping with the mobility challenge despite their relatively good performance when the mobility challenge is excluded while all the other challenges are kept.

6. CONCLUDING REMARKS AND FUTURE WORK

We have investigated in this paper how well contemporary image-set-based methods combined with fiducial-point-based features can be used for active authentication on smartphones. A dataset of 750 videos was collected over three sessions with different illumination conditions to capture the kind of variations that are likely to be present with mobile devices. An examination of the videos in the dataset revealed a unique combination of properties and challenges that is specific to smartphone face videos. We utilized the fact that the user's head is always close to the phone to increase the efficiency and reduce the false positives of the face detection phase. Although the compared state-of-the-art techniques perform relatively well when the enrollment and evaluation data come from the same session, the experiments indicate that they have difficulty addressing the variations in illumination and context that are likely to be present due to the mobility of the device.

One of the limitations of our study is that all the three sessions of any given user are collected in the same day. Therefore, the dataset misses appearance variations due to change in hair style, shaving, and/or introduction/removal of face-covering clothing such as scarves or hats. This does not limit the usefulness of the dataset since it captures a subset of the practically possible variations that has already been shown through experiments to be challenging to the state-of-the-art algorithms included in the comparison.

The benchmark presented in this paper motivates the development of better features and recognition algorithms that are invariant to the mobility challenge yet efficient to compute. Also the detection and classification of partial faces need further research to allow video clips with partial faces to be processed rather than incorrectly getting flagged them as face-empty.

7. REFERENCES

[1] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, "Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 136–148, Jan 2013.

[2] Tao Feng, Ziyi Liu, Kyeong-An Kwon, Weidong Shi, B. Carbunar, Yifei Jiang, and N. Nguyen, "Continuous mobile authentication using touchscreen gestures," in *IEEE Conference on Technologies for Homeland Security*, Nov 2012, pp. 451–456.

[3] M.O. Derawi, C. Nickel, P. Bours, and C. Busch, "Unobtrusive user-authentication on mobile phones using biometric gait recognition," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Oct 2010, pp. 306–311.

[4] Abena Primo, Vir V Phoha, Rajesh Kumar, and Abdul Serwadda, "Context-aware active authentication using smartphone accelerometer measurements," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014, pp. 98–105.

[5] Matthew A Turk and Alex P Pentland, "Face recognition using eigenfaces," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1991, pp. 586–591.

[6] Peter N. Belhumeur, João P Hespanha, and David Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 19, no. 7, pp. 711–720, 1997.

[7] Kamran Etemad and Rama Chellappa, "Discriminant analysis for recognition of human face images," *Journal of the Optical Society of America A*, vol. 14, pp. 1724–1733, 1997.

[8] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, no. 2, 2009.

[9] Hakan Cevikalp and Bill Triggs, "Face recognition based on image sets," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2567–2573.

[10] Yiqun Hu, Ajmal S Mian, and Robyn Owens, "Sparse approximated nearest points for image set classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 121–128.

[11] Yi-Chen Chen, Vishal M Patel, P Jonathon Phillips, and Rama Chellappa, "Dictionary-based face recognition from video," in *Computer Vision–ECCV 2012*, pp. 766–779. 2012.

[12] Enrique G Ortiz, Alan Wright, and Mubarak Shah, "Face recognition in movie trailers via mean sequence sparse representation-based classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3531–3538.

[13] Kuang-Chih Lee, Jeffrey Ho, Ming-Hsuan Yang, and David Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Computer Vision and Image Understanding (CVIU)*, vol. 99, no. 3, pp. 303–331, 2005.

[14] Kuang-Chih Lee and David Kriegman, "Online learning of probabilistic appearance manifolds for video-based recognition and tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1, pp. 852–859.

[15] P Jonathon Phillips, Patrick J Flynn, J Ross Beveridge, W Todd Scruggs, Alice J O'Toole, David Bolme, Kevin W Bowyer, Bruce A Draper, Geof H Givens, Yui Man Lui, et al., "Overview of the multiple biometrics grand challenge," in *International Conference on Advances in Biometrics*, 2009, pp. 705–714.

[16] Alice J OToole, P Jonathon Phillips, Samuel Weimer, Dana A Roark, Julianne Ayyad, Robert Barwick, and Joseph Dunlop, "Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach," *Vision Research*, vol. 51, no. 1, pp. 74–83, 2011.

[17] Chris McCool and Sébastien Marcel, "Mobio database for the icpr 2010 face and speech competition," Tech. Rep., Idiap, 2009.

[18] M Gunther, Artur Costa-Pazo, Changxing Ding, Elhocine Boutellaa, Giovani Chiachia, Honglei Zhang, Marcus de Assis Angeloni, Vitomir Struc, Elie Khoury, Esteban Vazquez-Fernandez, et al., "The 2013 face recognition evaluation in mobile environment," in *International Conference on Biometrics (ICB)*, 2013, pp. 1–7.

[19] Marco Pedersoli Luc Van Gool Markus Mathias, Rodrigo Benenson, "Face detection without bells and whistles," in *European Conference on Computer Vision (ECCV)*, 2014, vol. 4, pp. 720–735.

[20] Paul Viola and Michael J Jones, "Robust real-time face detection," *International journal of computer vision (IJCV)*, vol. 57, no. 2, pp. 137–154, 2004.

[21] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic, "Robust discriminative response map fitting with constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3444–3451.

[22] "http://ibug.doc.ic.ac.uk/resources/drmf-matlab-code-cvpr-2013/," .

[23] Kilian Q Weinberger and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research (JMLR)*, vol. 10, pp. 207–244, 2009.