Annotated Crowd Video Face Database

Tejas I. Dhamecha*, Priyanka Verma*, Mahek Shah*, Richa Singh, Mayank Vatsa
IIIT-Delhi, India
{tejasd,priyanka13100,mahek13106,rsingh,mayank}@iiitd.ac.in

Abstract

Research in face recognition under constrained environment has achieved an acceptable level of performance. However, there is a significant scope for improving face recognition capabilities in unconstrained environment including surveillance videos. Such videos are likely to record multiple people within the field of view. Face recognition in such a setting poses a set of challenges including unreliable face detection, multiple subjects performing different actions, low resolution, and sensor interoperability. In general, existing video face databases contain one subject in a video sequence. However, real world video sequences are more challenging and generally contain more than one person in a video. Therefore, in this paper, we provide an annotated crowd video face (ACVF-2014) database, along with face landmark information to encourage research in this important problem. The ACVF-2014 dataset contains 201 videos of 133 subjects where each video contains multiple subjects. We provide two distinct use-case scenarios, define their experimental protocols, and report baseline verification results using OpenBR and FaceVACS. The results show that both the baseline results do not yield more than 0.16 genuine accept rate @ 0.01 false accept rate. A software package is also developed to help researchers evaluate their systems using the defined protocols.

1. Introduction

Face recognition has been a research area for more than five decades now [7, 16, 22] and it has matured enough to be used in actual applications such as face tagging, mobile phone unlocking, and time-attendance. In controlled environment the state-of-the-art face recognition systems in controlled environment achieve up to 0.997 verification rate at 0.001 false accept rate (FAR) [3, 14]. However, as we move from constrained to unconstrained environments state-of-the-art performance reduces [14]. The unconstrained environment would include (but not limited to) acquisition using low cost devices, varying lighting conditions, minimum user co-operation, and presence of multiple subjects within the field of view. In recent years, researchers have been working on designing video-based face recognition algorithms [2, 4, 8, 20] to address some of these challenges.

Figure 1: A law enforcement application scenario where subjects are matched using surveillance footages only. Top row of the figure shows four frames/images from the Boston bombing case. The suspects (the subject in black hat and the subject in white hat) can be seen walking along with other subjects. The bottom row show the face regions of the suspects.

An efficient system that works in unconstrained environment is likely to be useful in multiple applications. One such important scenario is when both gallery (target) and probe (query) are obtained without requiring user cooperation. For instance, the Boston bombing case in which only the CCTV footage of suspects are available and we want to match these against other CCTV footages to identify the suspect's movement. As shown in Figure 1, gallery and probe images/videos are typically obtained from a surveillance footage that may contain multiple subjects. In order to facilitate law enforcement agencies, it is critical for face recognition research to attain impressive performance in the aforementioned application scenario [5]. Further, these applications also involve addressing emerging covariates [6] of low image quality, varying resolution, and sensor interoperability, along with traditional covariates of pose, illumination and expression [22] as well as age and weight varia-

* Equal contribution from student authors.

Table 1: Details of existing video face databases. The proposed dataset, ACVF-2014, records crowd (multiple subjects) in motion in every video.

Dataset	Description	# Subjects	# Videos
Face In Action [11]	passport checking scenario (constrained), single subject/video	180	6,470
YouTube Faces[20]	unconstrained, celebrity videos, single subject/video	1,595	3,425
PaSC [3]	unconstrained, single subject/video	265	2,802
ChokePoint [21]	unconstrained, fixed camera surveillance, single subject/frame	54	48
SN-Flip [1]	almost still subjects, multiple subjects/frame	190	28
ACVF-2014	unconstrained, hand held devices, multiple subjects/frame	133	201

(a) Device I

(b) Device II

(c) Device III

Figure 2: Sample frames from the ACVF-2014 database. Multiple people appear together in each video along with subjects appearing in indoor unconstrained environment. The videos are captured using three different devices with different sensors and resolutions. The videos are captured while subjects are walking through a passage or passing through doors.

tions [18].

Table 1 presents a summary of the existing video face datasets. Face-In-Action (FIA) [11] database was created with focus on a typical border-security-passport-checking scenario, thus expecting user cooperation. In 2011, Wolf *et al*. [20] created the YouTube Faces (YTF) database, which focuses on unconstrained face recognition. The dataset consists of celebrity videos collected from a famous video-sharing website Youtube[1]. It provides predefined protocol sets and current state-of-the-art results report around 90% accuracy with approximately 9% equal error rate (EER) [19]. Recently, the point and Shoot Challenge database (PaSC) [3] has taken the unconstrained face recognition environment to the next level. The PaSC database contains single subject videos captured using handheld and high definition devices. On the pre-defined protocol, the baseline results are up to 49% verification accuracy at 1% FAR whereas the best performance of up to 93.4% has been reported by Goswami *et al*. [12].

In FIA, PaSC, and YTF databases, every video sports only one subject. However, in real world unconstrained environment, this is a difficult constraint. Recently, Barr *et al*. released the SN-Flip [1] database where each video contains multiple subjects. However, all the subjects in this database are almost still, thus it may not be well suited to evaluate realistic crowd video matching scenario, i.e. multiple subjects performing some actions.

It is our assertion that there is a significant scope for improving face recognition performance in unconstrained environment, particularly in crowd video scenarios where enrollment videos/images are obtained in unconstrained settings. To encourage research in this important area, we have prepared a dataset consisting of 201 videos pertaining to 133 subjects, where each video contains multiple subjects. The key contributions of this paper are:

[1] www.youtube.com

Table 2: Details of the Annotated Crowd Video Face Database-2014.

Device (Resolution)	# Videos	# Frames	# Subjects	# Subjects/Video			# Faces			
				Min	Max	Avg	G.Truth	Automatic Detection	False Detects (Removed)	Final Detects (Used)
Device I (640×480)	115	16,704	120	1	14	2.8	22,635	13,973	4,415	9,558
Device II (2304×1296)	72	9,566	116	1	10	2.3	12,263	10,459	3,071	7,388
Device III (1920×1080)	14	1,741	20	1	4	2.1	2,563	3,309	1,267	2,042
Total	**201**	**28,011**	**133**	**1**	**14**	**2.6**	**37,461**	**27,741**	**8,753**	**18,988**

Table 3: Number of videos per subject in the ACVF dataset. For example there are 23 subjects appearing in exactly 2 videos.

# Subjects	44	23	19	12	39
# Videos	1	2	3	4	≥5

1. Annotated Crowd Video Face (ACVF) Database-2014 includes videos/frames along with landmarks of faces in each frame. 10 times random subsampling based cross validation protocol files and a MATLAB software package for evaluation is also included.

2. To establish the baseline, the results are reported using OpenBR [17] and a commercial-off-the-shelf system, FaceVACS [2]. The results are shown on two different experimental protocols.

We also plan to actively maintain a results webpage for streamlined comparative analysis on the database via: https://research.iiitd.edu.in/groups/iab/acvf.html.

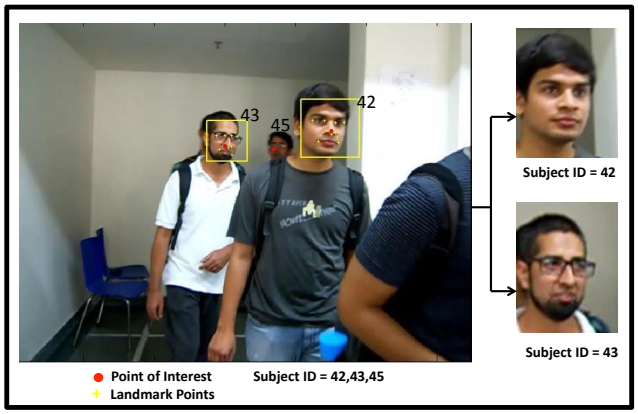


Figure 3: The annotation and face detection on an example frame. There are three POIs marked, where as the face detection algorithm detects two faces. POIs that are surround by each face-box are used to assign ground-truth subject IDs with each extracted faces. Also, there are some failures in detection cases, e.g. subject 45 is not detected in this example.

2. ACVF-2014 Dataset

The proposed ACVF-2014 database contains 201 videos (28,011 frames) of 133 subjects, captured at various locations, and each video contains up to 14 subjects. Consent for collecting these videos is taken from all the subjects. Some sample frames are shown in Figure 2. Typically, in all the videos, subjects appear in groups; therefore, almost all the video frames contain more than one subject. The videos are recorded using handheld devices without mounting on any tripod or similar structure. The dataset details are described below and a summary is provided in Table 2.

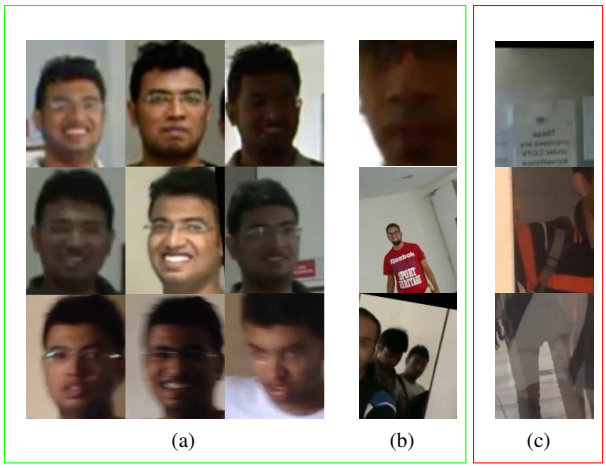
2.1. Device Details

The data is collected using three portable handheld devices having different resolutions. These devices are: Nikon Coolpix S570, Sony handycam DCR-DVD910E, and iPhone (4s and 5c). The three devices are referred to as Device I, Device II, and Device III respectively. Note that the device difference leads to varying quality of captured videos. The selection of these devices also introduces cross-sensor and cross-resolution covariates in the database.

2.2. Annotation, Face Detection, and Registration

Subject IDs along with a *point of interest* of all the faces present in a frame are manually annotated. Point of interest (POI) is a manually marked single point which is surrounded by the face box. We utilize the publicly available code of Everingham *et al*. [9] for face detection and provide the cropped faces of size 125×160. The face detection algorithm also finds nine landmarks points from the face region: two corners of both the eyes, two corners of lips, two corners of nose, and a nose tip. These nine landmark

[2] http://www.cognitec.com/facevacs-sdk.html



(a) (b) (c)

Figure 4: (a) Examples of accurately detected faces corresponding to each of the three devices. (b) sample of inaccurate face detection such as partial face and presence of extra non-face/background regions, and (c) shows examples of false detections which are discarded based on the POI annotations.

points are utilized to register a detected face with a canonical face frame. A subject ID is assigned to each extracted face image based on the POI. The procedure is illustrated in Figure 3. If no POI falls within a detected face rectangle, it is considered as incorrect/false face detection. It is possible that, even after POI based filtering, partial faces and faces with background information may be obtained (see Figure 4b). They may be considered inaccurate face detects. Figure 4 shows samples of detected and registered faces, inaccurately detected faces, and false detections. Due to the presence of covariates such as low resolution, blur, and nonuniform lighting, not all the faces are successfully detected. As mentioned in Table 2, out of the total manually marked 37,461 faces, only 27,741 faces are detected, out of which 8,753 face images are discard based on POI annotations (Figure 4c shows some failure cases). Thus, the remaining set of 18,988 faces is utilized in the experiments. Figure 5 illustrates the number of detected faces of each subject in each video along with the respective ground truth information[3]. Each registered output face image is named using the following convention.

$$DeviceName\_VideoID\_FrameNo\_SubjectID.jpg$$

Moreover, registered face images are provided in the `/Cropped/DeviceName/VideoID` directory for easier access. The example of image naming and directory structure is given in Figure 6.

[3]Cropped face images are provided in the dataset package; However, we encourage researchers to apply their own face detection algorithms.

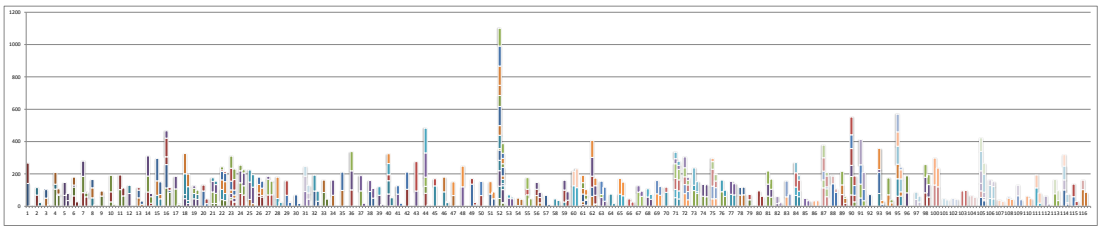
3. Application Scenarios and Experimental Protocols

As mentioned earlier, the ACVF dataset focuses on unconstrained face recognition (to be precise, face verification) with multiple subjects in a video/image. With these variations, there are two application scenarios in which this database can be utilized:

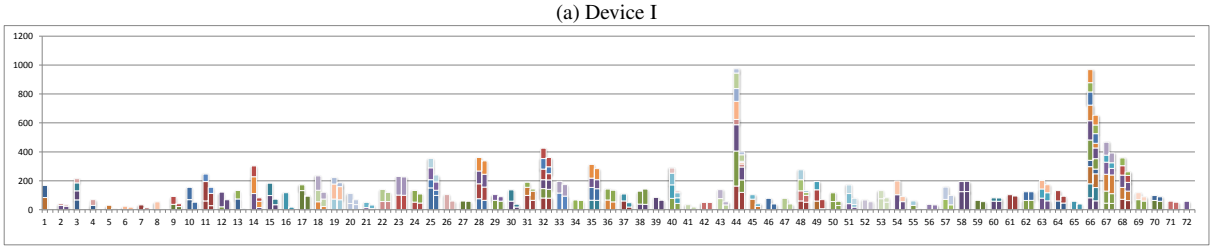
1. In Scenario I, the gallery set is defined in terms of a *set of videos*. Let the gallery set be defined as $\mathcal{G} = \{I_{v,f,n} | v \in \mathcal{V}\}$; where $\mathcal{V}$ is the set of video IDs selected to be the part of gallery and the n^{th} detected face image from a frame f of video v be denoted as $I_{v,f,n}$. This scenario has the following three different evaluation settings, each associated with a certain real world application:

   - **Frame-to-Frame Matching:** Scores are obtained by matching every face image (frame) in the probe set with every face image (frame) in the gallery set. The comparison of a probe video consisting of m face images and a gallery video consisting of n face images results in mn match scores.
   - **Video-to-Frame Matching:** The probe face frame is compared against every video in the gallery set. A set of scores is obtained by comparing a probe face frame with all the face frames in the gallery video. If the gallery video consists of q subjects, the set of scores are divided into q subsets, each corresponding to one subject. The scores within each subset are aggregated to obtain a match score between a probe face image (frame) and a gallery subject. Therefore, comparison of a probe video consisting of m face images (frames) against a gallery video consisting of q subjects, results in mq match scores.
   - **Video-to-Video Matching:** The probe video set is compared against the gallery video set. Each of the probe face images (frames) are compared with all the face images (frames) in the gallery video. For every video pair matching, the set of scores are aggregated such that a match score is obtained for every subject-pair comparison. Therefore, comparison of a probe video consisting of p subjects against a gallery video showing q subjects results in pq match scores.

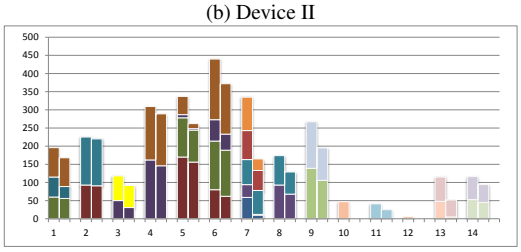
   Note that in all the three cases, the scores of one probe video comparison must not affect the scores of another probe video. For Scenario I, the videos for the gallery set are chosen such that every subject is present in at least one of the videos. The process of obtaining



(a) Device I



(b) Device II



(c) Device III

Figure 5: Representing the results of face detection. Two stacked bars are shown side-by-side for each video: first stacked bar represents the number of ground truth faces and the second staked bar represents the number of detected faces. The subparts of the bar (shown in different colors) represent each subject in the video. For example, video # 1 from Device III shows that there are three subjects (green, blue and orange) in the video. Note that the presence of more colors in one stacked bar translates to larger crowd (subjects).

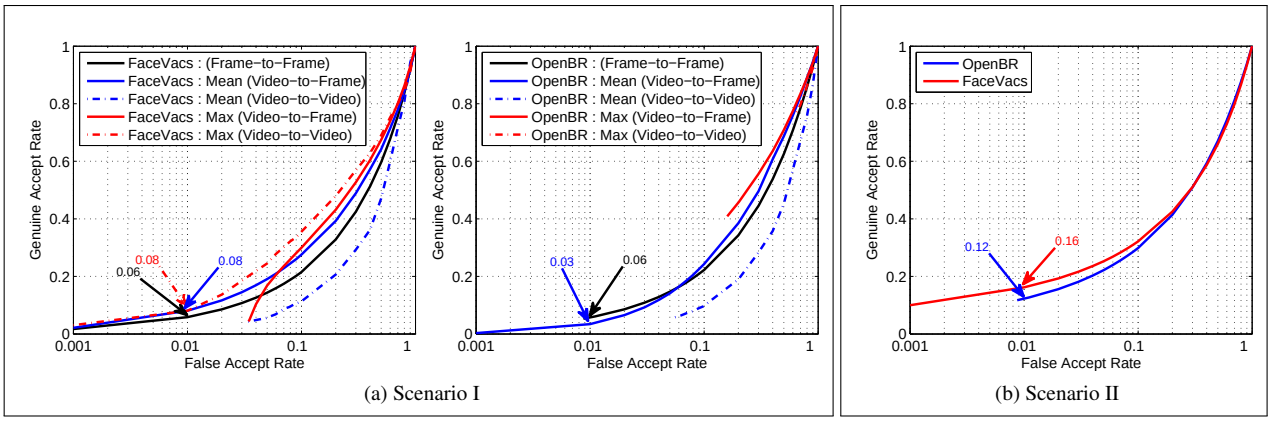
```
Cropped
├── Digicam
│   └── ...
├── IPhone
│   └── ...
└── Sonycam
    ├── Sony1
    │   └── ....
    └── Sony20
        ├── Sonycam_Sony20_001_115.jpg
        ├── Sonycam_Sony20_002_115.jpg
        └── ...
```

Figure 6: Directory structure of the cropped face images provided as part of dataset package.

gallery-probe split of videos is repeated 10 times to obtain the cross validation sets. The number of videos in the gallery set ranges between 61 to 71. These cross validation sets are included in the evaluation package.

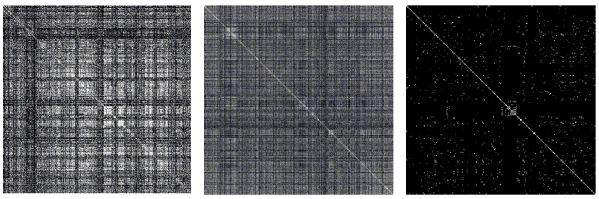
2. In Scenario II, the gallery set is defined in terms of a *set of images*; i.e. the video information is not considered, and images are referred using only indices. In this scenario, the gallery set is defined as $\mathcal{G} = \{I_k | k \in \mathcal{K}\}$; where I_k denotes the k^{th} image and $\mathcal{K}$ is the set of image indices selected to be a part of the gallery set. In this protocol, it is possible that both gallery and probe may be from the same video. This protocol helps to understand the performance of algorithms when face matching is required within a video, at different time stamps. 10 images per person are randomly selected to constitute the gallery set, while all the remaining images constitute the probe set. There are 8 subjects having less than 10 images and therefore, all the images pertaining to these subjects are included in the probe set. Thus, the gallery set contains 1250 (125 subjects, 10 images/subject) images, and the probe set contains 17,738 images.

It is important to note that the database is designed to evaluate face recognition systems and therefore, training data is not provided. Researchers may use any data (in any amount) from other sources but non-overlapping from the



(a) Scenario I (b) Scenario II

Figure 7: On the proposed ACVF-2014 database, ROC curves showcasing the verification performance of (a) FaceVACS (left) and OpenBR (right) for different settings of Scenario I, and (b) Scenario II. In Scenario II, since no frame or video associations are considered while generating the gallery probe splits, this scenario is close to still-to-still matching.



(a) OpenBR (b) FaceVACS (c) Ideal

Figure 8: Visualization of 18,988×18,988 similarity matrices obtained from (a) OpenBR and (b) FaceVACS. (c) shows the ideal similarity matrix for the given database. Darker pixels represent lower similarity between the corresponding gallery and probe image pair. All the three matrices are symmetric.

ACVF-2014 database to train their algorithms e.g. CMU-MultiPIE [13] and LFW [15]. This makes the evaluation completely non-overlapping and blind, which is the case with real world uncontrolled face recognition applications.

4. Baseline Results

Baseline evaluations have been performed using OpenBR [17] and FaceVACS (which is among the best commercial face recognition systems [14]). The face recognition module of OpenBR is based on Spectrally Sampled Structural Subspaces Features algorithm, also known as 4SF. For OpenBR, a built-in face detection module is used, whereas, for FaceVACS, eye coordinates are provided for each face image to ensure 100% enrollment in gallery. The verification performance is reported in terms of receiver operating characteristic (ROC) curve. The ROCs obtained for each cross-validation spit are combined into one curve using vertical averaging [10]. The results for Scenarios I and II are reported in Figures 7a and 7b, respectively. The key observations are:

- In both the scenarios, at 0.01 FAR, the best verification rate achieved is only 0.16. Further, many ROC curves start around 0.05 FAR which is likely to happen when the match score distribution does not have a long tail. This poor performance indicates the complexity of the problem as well as the limitation of the current systems.

- In both the scenarios, FaceVACS appears to perform slightly better than OpenBR. However, at higher FARs, the performance difference is not significant. It should be noted that eye annotation information is provided as an additional input to FaceVACS whereas OpenBR operates on loosely cropped faces from which it has to detect face region on its own.

- Score aggregation for video-to-video and video-to-frame matching is performed using two strategies: mean and max. Since both the systems provide similarity scores, the max strategy translates to selecting the scores corresponding to the best match. Both the systems suffer significantly in video-to-video matching using mean aggregation strategy and the best performance is observed with video-to-video matching with max aggregation strategy. This result underlines the importance of frame selection [12].

- At low FAR (≤ 0.01), Scenario II yields slightly better verification rate than Scenario I. In Scenario II, it is possible (and also likely) to have images of a subject from the same video in gallery as well as in probe.

Intuitively, they should be easier to match and such scores are leading to the minor improvement in performance.

- Figure 8 shows the similarity matrices (symmetric) of both the systems obtained by comparing all the detected faces with each other. The ideal similarity matrix is also shown, which has value 1 for all the genuine scores and 0 for all the impostor scores. The entropy of this matrix is very low whereas the entropies of the other two matrices are very high. This analysis substantiates the results obtained from ROC curves that a significant effort is required to achieve higher accuracies on the AVCF database.

5. Evaluation Package and Guideline

In the evaluation package we provide:

- Raw videos, detected faces images, and annotation information (POI and face landmark points detected using [9]),
- Protocol files and mask matrices, and
- MATLAB code for end-to-end evaluation.

The package is designed to make the overall evaluation process as easy as possible. To carry out the evaluation analysis, a 18,988×18,988 similarity matrix is required as input. Various evaluations, as discussed in this paper, can be performed from this similarity matrix and all the protocol files are provided as part of the package.

Acknowledgement

This work is supported by a grant from the Department of Electronics and Information Technology, India. T. I. Dhamecha is partly supported through TCS PhD research fellowship.

References

[1] J. R. Barr, L. A. Cament, K. W. Bowyer, and P. J. Flynn. Active clustering with ensembles for social structure extraction. In *IEEE WACV*, pages 969–976, 2014. 2

[2] L. Best-Rowden, B. Klare, J. Klontz, and A. K. Jain. Video-to-video face matching: Establishing a baseline for unconstrained face recognition. In *IEEE BTAS*, pages 1–8, 2013. 1

[3] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, et al. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE BTAS*, pages 1–8, 2013. 1, 2

[4] H. Bhatt, R. Singh, and M. Vatsa. On recognizing faces in videos using clustering-based re-ranking and fusion. *IEEE TIFS*, 9(7):1056–1068, 2014. 1

[5] H. Bhatt, R. Singh, M. Vatsa, and N. Ratha. Improving cross-resolution face matching using ensemble-based co-transfer learning. *IEEE TIP*, 23(12):5654–5669, 2014. 1

[6] H. S. Bhatt, R. Singh, and M. Vatsa. Covariates of face recognition. Technical Report IIITD-TR-2015-002, IIIT-Delhi, 2015. 1

[7] W. W. Bledsoe. The model method in facial recognition. *Panoramic Research Inc., Rep. PR1*, 15:47, 1966. 1

[8] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *ECCV*, pages 766–779. Springer, 2012. 1

[9] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in TV video. *IVC*, 27(5):545–559, 2009. 3, 7

[10] T. Fawcett. An introduction to ROC analysis. *PRL*, 27(8):861–874, 2006. 6

[11] R. Goh, L. Liu, X. Liu, and T. Chen. The CMU face in action (FIA) database. In *F&G*, pages 255–263. 2005. 2

[12] G. Goswami, R. Bhardwaj, R. Singh, and M. Vatsa. MDLFace: Memorability augmented deep learning for video face recognition. In *IEEE/IAPR IJCB*, pages 1–7, 2014. 2, 6

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *IVC*, 28(5):807–813, 2010. 6

[14] P. Grother, G. Quinn, and P. Phillips. Multiple biometric evaluation (MBE) 2010, report on the evaluation of 2D still-image face recognition algorithms. *NIST Interagency Report*, 7709, 2010. 1, 6

[15] 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. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 6

[16] A. K. Jain, P. Flynn, and A. A. Ross. *Handbook of biometrics*. Springer, 2007. 1

[17] J. C. Klontz, B. F. Klare, S. Klum, A. K. Jain, and M. J. Burge. Open source biometric recognition. In *IEEE BTAS*, pages 1–8, 2013. 3, 6

[18] M. Singh, S. Nagpal, R. Singh, and M. Vatsa. On recognizing face images with weight and age variations. *IEEE Access*, 2:822–830, 2014. 2

[19] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *IEEE CVPR*, pages 1701–1708, 2014. 2

[20] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE CVPR*, pages 529–534, 2011. 1, 2

[21] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *IEEE CVPR Workshops*, pages 81–88, 2011. 2

[22] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003. 1