Cross-Spectral Cross-Resolution Video Database for Face Recognition

Maneet Singh, Shruti Nagpal, Nikita Gupta, Sanchit Gupta, Soumyadeep Ghosh,
Richa Singh, and Mayank Vatsa
IIIT-Delhi, New Delhi, India
{maneets, shrutin, nikita13068, sanchit13088, soumyadeepg, rsingh, mayank}@iiitd.ac.in

Abstract

Advancing state of the art in face recognition and bridging the gap between laboratory and real-world scenarios require availability of challenging databases. One of the challenging applications in face recognition is surveillance, where unconstrained video data is captured both in day and night time (visible and near infrared) with multiple subjects in frames, which are matched with good quality gallery images. Due to the lack of an existing database for such a cross spectral cross resolution video-to-still face recognition application, this is still an open research problem. This paper presents a video database that can be utilized to benchmark face recognition algorithms addressing cross spectral cross resolution matching. The proposed Cross-Spectral Cross-Resolution Video dataset (CSCRV) contains videos pertaining to 160 subjects with an open-set protocol. We present baseline results with two commercial matchers for two experimental scenarios, where we observe very low performance of both the matchers. It is our assertion that this dataset can help researchers develop robust face recognition algorithms to handle real world surveillance scenarios.

1. Introduction

Advancements in the field of science and technology induces higher standards of living and improvement in the quality of human life. It enables individuals as well as organizations to be more aware of their surroundings by providing them with tools to develop a more secured environment. One such application for enhancing security is automated surveillance systems. Over the past few years such systems have gained immense popularity [16], primarily due to the convenient size of devices, ease of installation of the systems and effortless invigilation. Devices for unconstrained surveillance monitoring are being installed in public places such as markets, societies and community places. Though the hardware for unconstrained surveillance is readily available, automated identity recognition in such arduous settings still requires focused attention from the research community [13, 15].

In real-world surveillance scenarios, videos are captured in completely unconstrained environment, for both day and night time. For improved acquisition in night-time, surveillance devices often function in Near Infrared (NIR) domain, as opposed to visible spectrum for day-time acquisition [10, 11, 14, 19, 20]. Images of an individual captured at varying distances from the camera also leads to large variations in resolution of the face images obtained [5, 6]. These two problems combined together give rise to the problem of cross-spectral, cross-resolution face recognition. Coupled with varying pose, illumination, movement and presence of multiple subjects in a frame, unconstrained face recognition in surveillance videos poses major challenges for fully automated systems.

Recently, on 26th March 2016, in the unfortunate incident of Brussels Bombing, the suspects were identified using the surveillance feed obtained from the airport and tram stations. As shown in Figure 1, one such frame released by the security agencies was used to identify the suspects. The unconstrained nature of this problem can be observed in the extracted frame. The successful use of such feed to help law enforcement agencies reinforces the need to develop an effective and automated system for real world scenarios.

Figure 1. Frame released from the video footage of the Brussels Bombing (2016) suspect.

Currently, there exist very few publicly available video

Table 1. Existing publicly available video datasets for face recognition

S.No.	Database	Number of Subjects	Number of Videos	Multiple Subjects in frame	Spectrum
1.	Face in Action (2005) [9]	180	6,470	No	VIS
2.	YouTube Faces (2011) [17]	1,595	3,425	No	VIS
3.	ChokePoint (2011) [18]	54	48	No	VIS
5.	PaSC (2013) [4]	265	2,802	No	VIS
6.	SN-Flip (2014) [3]	190	28	Yes	VIS
4.	McGillFaces (2015) [7]	60	60	No	VIS
7.	ACVF (2015) [8]	133	201	Yes	VIS
8.	**CSCRV (proposed)**	**160**	**193**	**Yes**	**VIS and NIR**

datasets for face recognition (see Table 1). Out of those given in Table 1, only SN-Flip [3] and ACVF [8] datasets contain videos having multiple subjects in one frame and all the existing databases are in visible spectrum. Videos belonging to SN-Flip dataset contain almost-still subjects with very less motion, while videos provided by ACVF are captured using hand-held cameras. To the best of our knowledge[1], none of the existing publicly available datasets provide videos in both visible and NIR spectrums captured at varying distances in unconstrained environment. In this research, we present the Cross-Spectral, Cross-Resolution Video (CSCRV) dataset, which consists of 193 videos of 160 subjects captured during the day (visible spectrum) and night (NIR spectrum). All videos capture the subjects in an unconstrained settings, walking from a distance of 10m, individually or in groups of two or more. In order to make the dataset more challenging and simulate real world scenarios, an open-set protocol is also presented for face identification. We perform baseline experiments using two commercial matchers and results show the challenging nature of the database. Section 2 provides details about the proposed dataset, whereas Section 3 and Section 4 discuss the experimental protocol and baseline results. This is followed by a section on the conclusions and practical applications of this dataset.

2. Dataset Description

The proposed dataset contains 193 videos of 160 subjects, where each video comprises of at least one and at most three subjects. Videos are captured at different locations and consists of both day and night time videos. Videos captured during the day-time are in the visible spectrum, whereas, night-time videos are captured in the NIR spectrum. Out of 193 videos, 98 videos are captured during the day-time while the remaining 95 are captured during the night-time. Figure 2 presents some sample frames from the dataset for two subjects.

[1]While there are a few cross-spectral face image databases, our focus is on video face recognition and therefore, we are not discussing still image databases.

2.1. Data Acquisition Setup

For data acquisition, a camera is mounted on a tripod stand and videos of subjects walking from a distance of $\mathbf{10m}$ are captured. For day-time videos, other than natural sunlight, no extra source of illumination is used. Visible pass filter is placed on the lens to ensure that only visible spectrum data is captured. For night-time videos, an NIR illuminator is placed behind the camera. A visible cut filter is placed on the lens which ensures that videos are captured in the NIR spectrum. The lighting from natural/artificial sources was left as it is. Figure 3 presents the setup used for data acquisition.

2.2. Devices Used

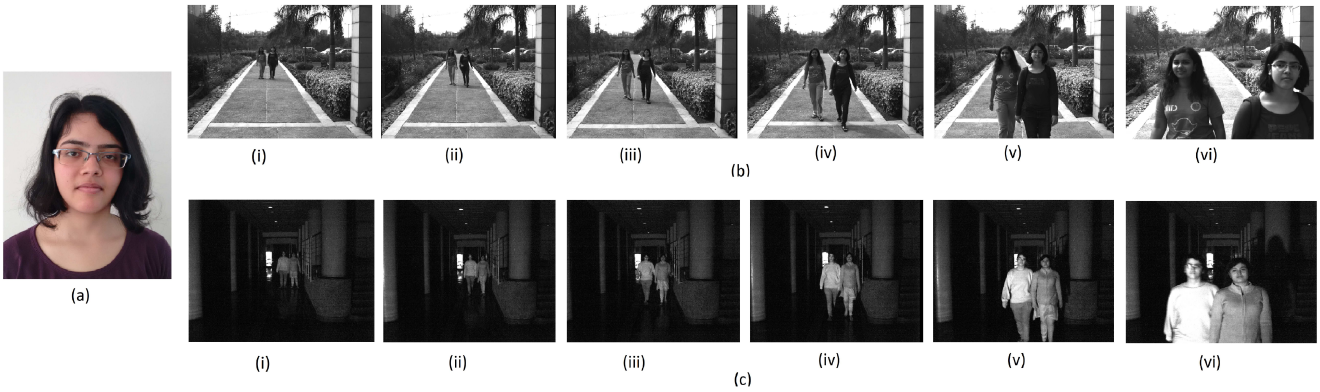
The following devices are used for data acquisition:

- GO-5000M-USB camera acquires monochromatic data in visible and NIR spectrum. The frames recorded are of 2560x2048 resolution in full 5-mega-pixel output. For better acquisition, auto gain and exposure of the camera are set to continuous and the frame rate is fixed to 20 frames per second. No compression is applied for storage of videos and a tripod stand is used to mount the camera, thereby ensuring stability during acquisition. These devices are consistently used in day-time as well as night-time data acquisition.
- For night-time, Advanced Illumination RL113-850IC (NIR) illuminator is used. It is used at 100% intensity to enhance quality of captured data.
- Still images are captured using smart phones having a resolution of at least 8 mega-pixel.

2.3. Volunteers and Dataset Statistics

The entire dataset consists of 87 male and 73 female subjects aged between 18-27 years. Out of 160 subjects, 125 subjects have at least one day-time video, one night-time video, and three high resolution still images. For a particular subject, all videos are captured on different days, thereby resulting in multiple sessions. 61 videos have one subject, while the remaining 132 videos have two or more subjects.

(A) Sample frames of Subject A



(B) Sample frames of Subject B

Figure 2. Sample frames from two videos of the CSCRV dataset. (a) corresponds to the high resolution still image of the subject, (b) represent some frames from the subject's day-time video and (c) represents some frames from the subject's night-time video.

Table 2. Details about the proposed CSCRV Dataset

Time of Day	Spectrum	Videos	Subjects	Mean Duration (secs)	Min. Duration (secs)	Number of Frames	Number of Detected Faces
Day	Visible	98	148	10.63	7.00	20,859	33,696
Night	NIR	95	158	10.57	7.00	20,091	34,714

Average length of the videos for the entire dataset is 10.60 seconds. The dataset contains 40,950 frames and 68,410 detected faces. Further details about the dataset are given in Table 2.

2.4. Nomenclature and Data Distribution

All videos are named in the following format: *'Time_LocationID_VideoID_SubjectID1_..._SubjectIDn'* Here, time refers to the time of the day the video was captured and may take two values, N or D. LocationID corresponds to the location at which the video was captured. It can take one of the four values: $S1$, $S2$, $S3$ or $S4$ ($S1$ and $S2$ refer to day-time locations, while $S3$ and $S4$ refer to night-time locations). VideoID corresponds to a unique ID given to each video of a location and SubjectID corresponds to a unique ID given to each subject. For example, consider the video name $N\_S4\_V28\_67\_0$, where N corresponds to a night-time video and $S4$ denotes that the video was captured in the fourth location. V28 denotes that video's unique ID and the remaining number(s) denote the subject IDs which are present in the video. Subject ID 0 corresponds to subjects belonging to the open-set. This nomenclature ensures that every video is given a unique and informative name. Each subject has three still images which have been named as SubjectID_1, SubjectID_2 and SubjectID_3 for each subject.

The dataset also includes annotated frames containing a bounding box for every face in each frame (total 68410 faces), following the nomenclature described above. Along with the loose cropped face images, each subject's three

Figure 3. Data acquisition setup: (a) represents subject at a distance of 10m from the camera, (b) shows the subject at a distance of 7m and in (c) the subject is at a distance of 1m.

high resolution still images are also part of the release. A small section of non-overlapping videos acquired under the same setup are also provided as a training set for learning-based experiments.

3. Experimental Protocol

Since CSCRV dataset captures both cross-distance and cross-spectral nature of the real world surveillance scenarios, two protocols (Cross-Distance and Cross-Distance Cross-Spectral) are proposed. In both protocols, high resolution still images of subjects are considered as *gallery*, while all videos are considered as *probes* (i.e. Still-to-Video matching). Out of 160 subjects, 35 are open-set subjects, i.e. their gallery images are not available. For each frame in all videos, a loose bounding box is obtained by tracking the first occurrence of every subject. This is done via the KLT algorithm [12] in MATLAB. Two commercial off-the-shelf systems (COTS), FaceVACS [1] and Luxand [2], are used to obtain baseline identification results on the proposed dataset for the two protocols explained below.

3.1. Scenario I: Cross-Distance Still-to-Video Face Matching

In the first scenario, visible spectrum videos are provided as probes which are to be matched with visible spectrum still images (gallery). A total of 125 subjects are enrolled in the gallery and 98 videos of 148 subjects are provided as probes, thereby resulting in an open-set protocol. On this, two baseline experiments are performed:

- **Experiment 1:** The entire video is considered as the probe and identification of subjects in each video is performed.
- **Experiment 2:** In the second experiment, results are reported on the basis of the distance of the subject from the camera. The videos are divided into three segments: $\mathbf{0-4m}$, $\mathbf{4-7m}$ and $\mathbf{7-10m}$. Recognition results for all three categories are computed individually.

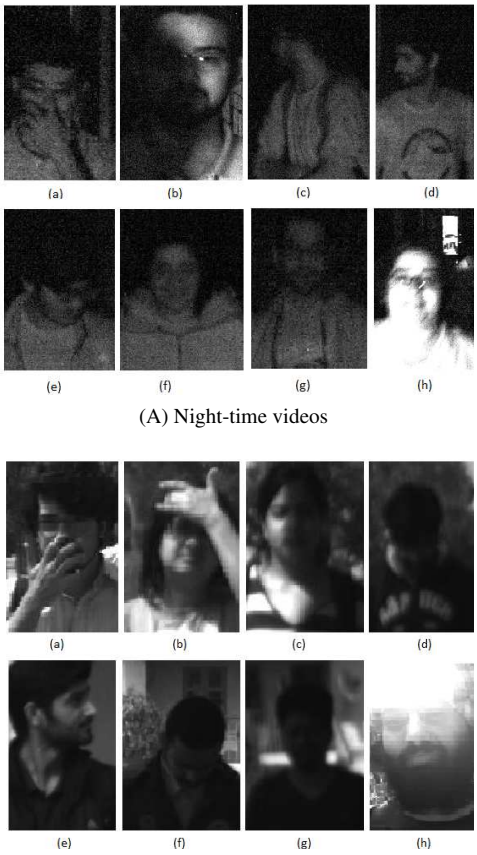
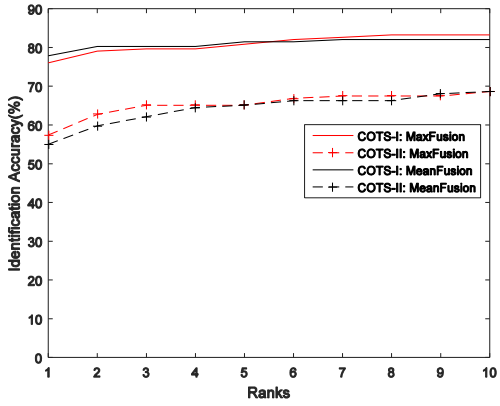


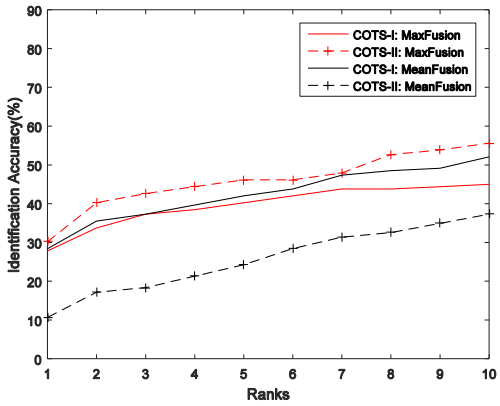
Figure 4. Sample challenges present in CSCRV dataset displaying covariates such as occlusion, shadow, pose variations, very low resolution and varying illumination.

3.2. Scenario II: Cross-Distance Cross-Spectral Still-to-Video Face Matching

In this scenario, NIR spectrum videos (night-time) are provided as probes which are to be matched with visible spectrum gallery images. A total of 125 subjects are enrolled in the gallery and 95 videos corresponding to 158 subjects are provided as probes. Similar to the previous ex-



(A) Cross-Resolution (Day)



(B) Cross-Resolution Cross-Spectrum (Night)

Figure 5. Cumulative Match Characteristics (CMC) curves obtained for experimental setup where the entire video was provided as probe, to be matched with a still high resolution gallery.

periments, two set of recognition results are reported:

- **Experiment 1:** In the first set of experiments, the entire video is considered as a probe and identification is performed on the entire data.
- **Experiment 2:** In the second experiment, videos are divided on the basis of the subject's distance from the camera. The videos are divided into three segments: $\mathbf{0-4m}$, $\mathbf{4-7m}$ and $\mathbf{7-10m}$. Recognition results for all three categories are computed individually.

Results corresponding to both the scenarios are given in the following section.

4. Results

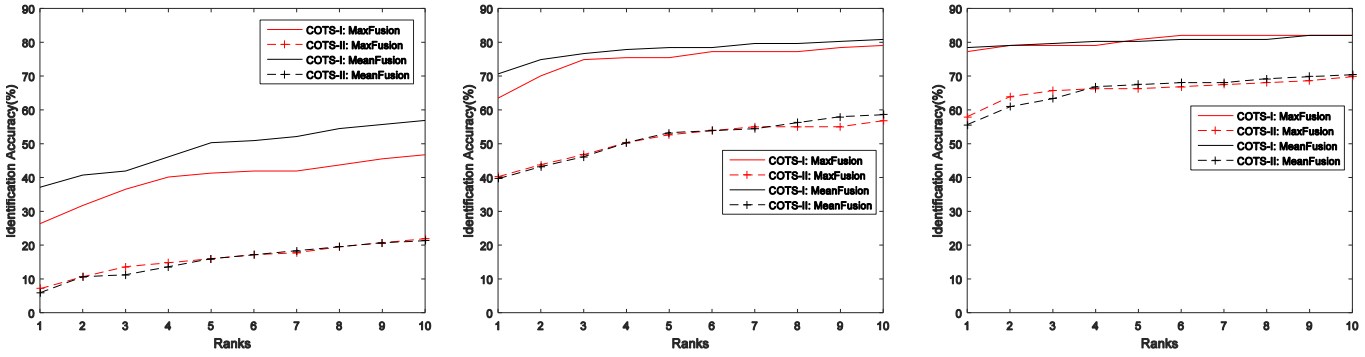
Figures 5 and 6 show the Cumulative Match Characteristics (CMCs) for experiments mentioned in Section 3. Two COTS, FaceVACS (COTS-I) and Luxand (COTS-II), are used for computing the baseline results. Table 3 shows the rank-1 identification accuracies for all the experiments. Since the protocols are for still-to-video matching, for a video, scores obtained for each frame are combined via (i) mean-score fusion and (ii) max-score fusion. Key results of the experiments are given below:

- In **cross-distance experimental protocol**, where the probes consist of visible spectrum videos and gallery consists of high resolution visible spectrum images, a rank-1 identification accuracy of 77.8% is obtained using COTS-I with mean-score fusion. At the same time, COTS-II results in a rank-1 identification accuracy of 57.4%.
- In **cross-distance cross-spectral experimental protocol**, where the probes are NIR spectrum videos and gallery consists of high resolution visible spectrum images, COTS-II gives a maximum rank-1 accuracy of 30.1% using max-score fusion. COTS-I yields a maximum accuracy of 28.4% using mean-score fusion.
- In **cross-distance experimental protocol**, when the subject is at a distance of $\mathbf{7-10m}$, COTS-II results in rank-1 identification accuracy of 7.1% by applying max-fusion on scores. COTS-I reports an accuracy of 37.1%. For the same setup, **cross-distance cross-spectrum** protocol results in a rank-1 accuracy of 1.8% using COTS-I and 1.2% using COTS-II. This set of experiments, where the subject is at the maximum distance from the camera, emphasizes upon the need to focus on cross-distance cross-spectral face matching in unconstrained surveillance scenarios.
- When the subjects are at a distance of $\mathbf{4-7m}$, slight improvement in rank-1 accuracies is observed. Middle column of Figure 6 presents the cumulative match characteristics curves for the distance of $\mathbf{4-7m}$.
- For the setup where the subjects are closest to the camera, $(\mathbf{0-4m})$, significant improvement is seen in rank-1 accuracies, as opposed to the setup where the subject is at the maximum distance. For **cross-spectral cross-distance protocol**, COTS-I and COTS-II report an accuracy of 33.7% and 30.2% respectively. When the probe videos contain visible spectrum videos, COTS-I reports a rank-1 accuracy of 78.4%, while COTS-II reports a maximum rank-1 accuracy of 57.9%. These results further justify the claim that face recognition at a distance is still a challenging problem and demands focused attention, especially for cross-spectral scenarios.

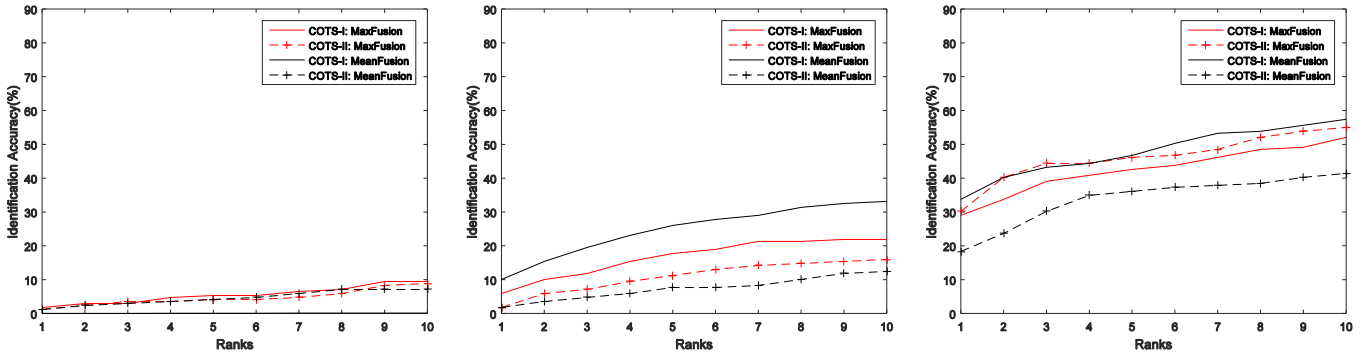
From Table 3, it can be seen that there is a considerable margin between the same spectral and different spectral face recognition performance. The results reported for the

Table 3. Rank-1 accuracies (%) obtained for four sets of experiments: entire videos as probe and distance based videos as probes ($\mathbf{7-10m}$, $\mathbf{4-7m}$ and $\mathbf{0-4m}$) using COTS-I and COTS-II. The results tabulated below correspond to mean based score fusion of frames.

Rank-1 Identification Accuracy (%)								
	Cross-Resolution (Day)				**Cross-Resolution Cross-Spectrum (Night)**			
	Entire Video	$\mathbf{7-10m}$	$\mathbf{4-7m}$	$\mathbf{0-4m}$	Entire Video	$\mathbf{7-10m}$	$\mathbf{4-7m}$	$\mathbf{0-4m}$
COTS-I	77.8	37.1	70.6	78.4	28.4	1.8	10.0	33.7
COTS-II	55.0	5.9	39.6	55.6	10.6	1.2	1.8	18.3



(A) Cross-Distance Face Recognition



(B) Cross-Spectral Cross-Distance Face Recognition

Figure 6. CMC curves for experiments performed on Visible and NIR videos of CSCRV Dataset. Image-to-video matching is performed, where gallery constitutes high resolution still images and probe constitutes the acquired videos. (A) corresponds to results obtained for day-time videos and (B) corresponds to the night-time videos. Each column represents distance based results, i.e. when the subject is at varying distances from the camera. Left to right: $\mathbf{7-10m}$, $\mathbf{4-7m}$ and $\mathbf{0-4m}$.

above mentioned protocols depict the challenging nature of the proposed dataset. To further motivate the use of the proposed dataset, Figure 4 gives some sample face images from the videos with challenging covariates. While we focused on still-to-video face matching, video-to-video matching can also be performed for cross-distance and cross-spectral scenarios.

5. Conclusion

In modern world, automating face recognition for 24 hour surveillance encompasses two major research hurdles, namely cross-distance (cross-resolution) and cross-spectral (VIS-NIR) face recognition. This work presents the Cross-Resolution Cross-Spectral Video (CSCRV) database which provides a platform to work on the challenging task of unconstrained face recognition in both cross-distance and cross-spectral scenarios. Experimental protocols are defined and baseline results obtained by commercial matchers illustrate the complex nature of the problem. To promote further research, this database will be made available to the research community.

References

[1] Facevacs. http://www.cognitec.com/.

[2] Luxand. https://www.luxand.com.

[3] J. R. Barr, L. A. Cament, K. W. Bowyer, and P. J. Flynn. Active clustering with ensembles for social structure extraction. In *Winter Conference on Applications of Computer Vision*, pages 969–976, 2014.

[4] J. R. Beveridge, J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, 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 *Biometrics: Theory Applications and Systems*, pages 1–8, 2013.

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

[6] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer. Pose-robust recognition of low-resolution face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):3037–3049, 2013.

[7] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel. Hierarchical temporal graphical model for head pose estimation and subsequent attribute classification in real-world videos. *Computer Vision and Image Understanding*, 136:128–145, 2015.

[8] T. I. Dhamecha, P. Verma, M. Shah, R. Singh, and M. Vatsa. Annotated crowd video face database. In *International Conference on Biometrics*, pages 106–112, 2015.

[9] R. Goh, L. Liu, X. Liu, and T. Chen. The cmu face in action (fia) database. In *Analysis and Modelling of Faces and Gestures*, pages 255–263. 2005.

[10] M. Grgic, K. Delac, and S. Grgic. Scface – surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, 2009.

[11] F. Juefei-Xu, D. K. Pal, and M. Savvides. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *Computer Vision and Pattern Recognition Workshops*, pages 141–150, 2015.

[12] Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-backward error: Automatic detection of tracking failures. In *International Conference on Pattern Recognition*, pages 2756–2759, 2010.

[13] D. Kang, H. Han, A. K. Jain, and S.-W. Lee. Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching. *Pattern Recognition*, 47(12):3750 – 3766, 2014.

[14] S. Z. Li, D. Yi, Z. Lei, and S. Liao. The casia nir-vis 2.0 face database. In *Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013.

[15] S. P. Mudunuri and S. Biswas. Low resolution face recognition across variations in pose and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):1034–1040, 2016.

[16] J. Pagliery. Fbi launches a face recognition system, 2014. [Online; posted 16-September-2014].

[17] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition*, pages 529–534, 2011.

[18] 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 *Computer Vision and Pattern Recognition Workshops*, pages 74–81, 2011.

[19] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li. Face matching between near infrared and visible light images. In *International Conference on Advances in Biometrics*, pages 523–530, 2007.

[20] J. Y. Zhu, W. S. Zheng, J. H. Lai, and S. Z. Li. Matching nir face to vis face using transduction. *IEEE Transactions on Information Forensics and Security*, 9(3):501–514, 2014.