FaceSurv: A Benchmark Video Dataset for Face Detection and Recognition Across Spectra and Resolutions

Sanchit Gupta, Nikita Gupta, Soumyadeep Ghosh, Maneet Singh, Shruti Nagpal, Mayank Vatsa, and Richa Singh
IIIT-Delhi, India

***Abstract*— Existing face recognition algorithms achieve high recognition performance for frontal face images with good illumination and close proximity to the imaging device. However, most of the existing algorithms fail to perform equally well in surveillance scenarios, where videos are captured across varying resolutions and spectra. In surveillance settings, cameras are usually placed far away from the subjects, thereby resulting in variations across pose, illumination, occlusion, and resolution. Current video datasets used for face recognition are often captured in constrained environments, and thus fail to simulate the real world scenarios. In this paper, we present the FaceSurv database featuring 252 subjects in 460 videos. The proposed dataset contains over 142K face images, spread across videos captured in both visible and near-infrared spectra. Each video contains a group of individuals walking from $36ft$ towards the imaging device, offering a plethora of challenges common to surveillance settings. Benchmark experimental protocol and baseline results have been reported with state-of-the-art algorithms for face detection and recognition. It is our assertion that the availability of such a challenging database will facilitate the development of robust face recognition systems relevant to real world surveillance scenarios.**

I. Introduction

Several instances of terrorism and public disorder has led to the widespread utility of automated surveillance systems [19]. Most public spaces including airports, railway stations, government buildings, and bus terminals are equipped with CCTV cameras in order to monitor daily movements. The availability of sophisticated and cost effective surveillance devices along with the ease of installation and invigilation has provided impetus to the video surveillance infrastructure at large. Such surveillance cameras often operate 24×7 (day and night) and are installed such that there is large standoff distance between the subjects and the camera, resulting in large variability of image quality, including low resolution and cross-spectrum data acquisition [4], [5]. Surveillance cameras often operate in the visible spectrum during the day-time, and switch to the Near Infrared (NIR) spectrum during the night-time [9], [11], [15], [29], [32]. In addition to that, surveillance systems do not require user co-operation, and generally capture data in completely unconstrained settings. In such settings, the captured face images are often partially occluded due to hand movement, facial hair, or accessories such as sunglasses or hats. Depending upon the location of the illumination source, shadows are also commonly observed in such unconstrained scenarios. In law enforcement applications, a recorded surveillance video might be matched against a pre-acquired database of individuals. Generally, the pre-acquired database contains high resolution, well illuminated face images, which are to be matched with the low quality, low resolution videos obtained from the surveillance systems. Along with the variation in spectrum (for night-time data captured in NIR spectrum), this results in the challenging problem of cross spectral cross resolution face recognition. Fig. 1 illustrates the problem of face recognition with variations due to resolution and spectrum. Challenges such as variations in pose and illumination, occlusions, and movement of multiple subjects in frames makes the problem much more challenging (Fig. 2). Though there has been significant improvement in the hardware technology over the past years, automated face recognition in surveillance settings still remains an open and challenging research problem [14], [18].

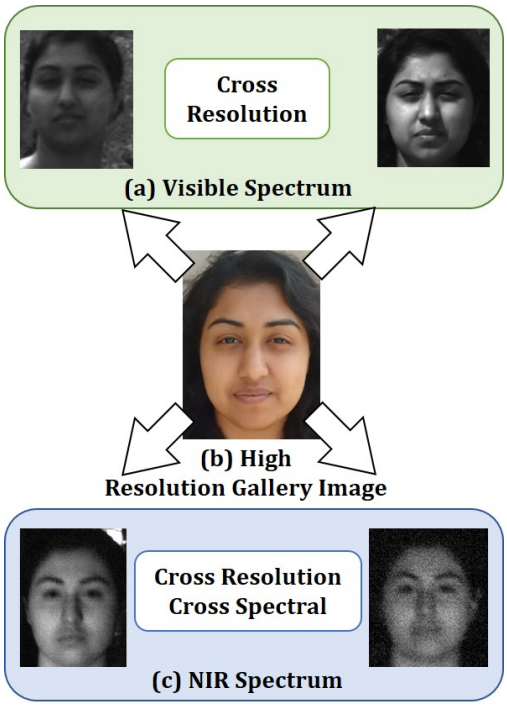


Fig. 1: The proposed FaceSurv dataset contains videos captured in two spectra: visible and NIR, along with high resolution gallery images. The proposed dataset enables matching in cross spectral, and cross spectral cross resolution scenarios.

The challenge of matching across spectra and resolutions is further accentuated due to the limited availability of large datasets. Table I presents a brief review of the publicly

Fig. 2: Challenges of the proposed FaceSurv dataset: (a) variations in pose, (b) occlusion of faces due to variants like hair, hands, caps, spectacles, and shadow, (c) variation in illumination and distance for a subject's sample video.

TABLE I: Literature review of video databases for face recognition.

Database	Spectrum	Subjects per Video	No. of	
			Subjects	Videos
Face in Action [8]	VIS	Single	180	6470
YouTube Faces [25]	VIS	Single	1595	3425
ChokePoint [26]	VIS	Single	54	48
PaSC [3]	VIS	Single	265	2802
SN-Flip [2]	VIS	Multiple	190	28
McGillFaces [6]	VIS	Single	60	60
CrowdFaceDB [7]	VIS	Multiple	257	385
CSCRV [22]	VIS&NIR	Multiple	160	193
IJB-S [13]	VIS	Multiple	202	350*
Proposed FaceSurv	**VIS&NIR**	**Multiple**	**252**	**460**

*The dataset contains 350 surveillance videos and 202 enrollment videos.

available datasets for addressing the problem at hand. It can be observed that only CrowdFaceDB [7], SN-Flip [2], CSCRV [22], and IJB-S [13] datasets have multiple subjects in video frames. However, videos in SN-Flip dataset contain almost stationary subjects undergoing very less movement, while videos in CrowdFaceDb are captured using hand-held devices. Moreover, almost all the datasets contain videos captured in visible spectrum only, without much variation of the subject distance from the camera. To the best of our knowledge, only the CSCRV dataset [22] contains videos captured across the two spectra (visible and NIR), having multiple subjects per video.

This research extends the existing literature and presents the proposed FaceSurv dataset which contains videos across two spectra and resolutions. The proposed dataset contains 460 videos of 252 subjects captured during the day (visible spectrum) and night (NIR spectrum). All videos capture the subjects (individually or in groups) in unconstrained settings, at a varying distance of 1*ft* to 36*ft*. Experimental protocols and baseline results have been provided for face detection and recognition. Face detection results are reported with the current state of the art face detection algorithms namely Viola-Jones [24], Fast Face Detector [16], Tiny Face Detector [10], and Single Shot Scale-Invariant Face Detector [31]. Experimental protocols and baseline results have also been reported for face recognition, where two experimental protocols have been presented: *Cross-Resolution (CR)* face recognition i.e. VIS (high resolution) to VIS (low resolution) matching and *Cross-Spectral Cross-Resolution (CS-CR)* face recognition i.e. NIR (low resolution) to VIS (high resolution) matching. Results have been reported with two commercial off-the-shelf systems (COTS), COTS-A [1] and Verilook (COTS-B) [1], along with deep learning feature extractors: LightCNN [27] and VGGFace [20]. In order to facilitate research in this direction, the proposed dataset will be released to the research community, along with the experimental protocols.

II. Proposed FaceSurv Dataset

The proposed FaceSurv database contains 460 videos of 252 subjects in the age range of 18-60 years, including some open-set subjects. Each video contains 1-4 subjects walking from a distance of $36ft$ towards the camera. The subjects were asked to walk freely in an unconstrained manner, without any restriction on pose, occlusion, or body movement. Videos have been captured at four locations: three outdoor and one indoor; during the day and night time, across multiple sessions. As shown in Table II, out of the 460 videos, 234 are captured during the day time, in the visible spectrum, while the remaining 226 videos are captured during the night time, in the near infrared (NIR) spectrum. This results in over 46000 frames captured in the visible spectrum, and over 44000 frames captured in the NIR spectrum. Table III presents the location based statistics of the proposed dataset. Each location has over 100 videos of over 150 subjects. Since data is captured across multiple sessions and locations, one subject can occur in multiple videos in the dataset, this results in a total of 721 subject instances across the dataset.

In order to simulate a real world law enforcement scenario, where high resolution mugshot images are matched with a low resolution input, the proposed dataset also contains three high resolution *gallery* images for every subject. Fig. 3 presents the high resolution gallery images along with sample

[1]Name of the commercial matcher is suppressed due to license restrictions, however, it is one of the top performing matchers in NIST evaluation.

Fig. 3: (1-a) and (2-a) correspond to the high resolution gallery images, and (1-b) and (2-b) present sample annotated faces captured in the visible and NIR spectrum, respectively, of the proposed FaceSurv dataset. Subjects are asked to walk from $36ft$ towards the camera, resulting in face images with varying resolutions.

TABLE II: Statistics of the proposed FaceSurv dataset.

Spectrum	Total Videos	Subject Instances	Average Duration	Frame Count	Total Faces
VIS (Day)	234	362	10.04 s	46,995	71,653
NIR (Night)	226	359	9.88 s	44,651	70,463

frames from two videos. In order to facilitate research in the area of face detection and recognition on the proposed dataset, ground truth face regions have also been computed by a combination of manual and automated annotations. Bounding box of the face region is manually annotated in the first frame, which is then tracked using the KLT tracker [12]. In case of incorrect tracking, the tracker is reset by manual intervention. Effort has been made to ensure that all the tracked regions are faces and neither is any face missed nor any non-face region included. In total, the proposed FaceSurv dataset contains over 140K face images.

A. Data Acquisition

As mentioned previously, the proposed dataset contains videos of subjects walking from a distance of $36ft$ towards a high resolution camera mounted on a tripod stand. In the day time, data is captured in sunlight, eliminating the need of any extra source of illumination. For data captured during nighttime, an NIR illuminator is used to capture data in the NIR spectrum. Since the camera is capable of collecting data in both visible and NIR spectrum, videos are collected during the day with a VIS-pass filter, while a VIS-cut filter is applied on the camera during the night time. Details regarding the devices used for acquisition are as follows:

1) **A 5 Megapixel USB camera** with Nikon DX VR 18-300mm lens is used to record the 460 surveillance videos in both day time and night time. The resolution of the acquired videos is 2560×2048, and they are monochromatic in nature. The videos are recorded at 20 frames per second using the Jai Control Tool Software. Auto-gain and exposure settings are set to continuous during acquisition for better perceptibility. The camera is mounted on a tripod stand for ensuring proper stability of the camera while data collection.
2) **Advanced NIR Illuminator** is used during the acquisition of near-infrared data for the night recordings. It is placed behind the camera facing the subject with an intensity of 100%.
3) **Smartphone** with at least 8 mega-pixel camera resolution is used for capturing the high resolution gallery images in a constrained environment, with good illumination, and minimal pose variations.

TABLE III: Location-wise statistics of the proposed dataset.

S.No.	Session Description	Number of Videos	Number of Subjects
1.	Location-1	124	184
2.	Location-2	110	168
3.	Location-3	112	170
4.	Location-4	114	173
5.	**Total Subjects**	**252**	
6.	**Total Videos**	**460**	
7.	**Total Instances**	**721**	

B. Data Distribution and Nomenclature

In order to facilitate research in the challenging area of cross-resolution cross-spectral recognition, the proposed dataset will be released for research purposes. All the videos are named in such a manner that they are uniquely identifiable, and provide information about the location and spectrum in which the video was captured, and the subjects present in it. Each video is named in the following format: *'T_LID_VID_SId$_1$...SId$_n$'*. Here, *T* corresponds to the time of the day when data was captured, and can have two values: either *D* (day) or *N* (night). The second component of the name, *LID*, is the identifier of the location where the video was captured. Since there are four locations in the proposed dataset, *LID* can take one of {S1, S2, S3, S4}. Here, S1, S2, and S4 refer to the three outdoor locations, while S3 refers to an indoor location. The third component of the video name refers to the *VID*, which is a unique identifier given to each video captured at a specific location. The final components of the video name correspond to the unique subject identifiers which are present in the video. If a subject ID is 0, it corresponds to an open-set subject. For example, a video of the proposed dataset can be named *'N_S4_V87_71_126_140'*. This refers to a video which was captured in the night time (NIR spectrum), at the fourth location, has a video identifier number 87, and contains subject IDs 71, 126, and 140.

Along with the videos, three high resolution gallery images are also provided for each subject which is not in the open-set. The gallery images are named as: *SID_1*, *SID_2* and *SID_3*. The proposed dataset also contains loosely bounded annotated faces for each video. As mentioned previously, the face region for each subject at each frame in a given video is annotated and stored. The nomenclature of the annotations are as follows: *T_LID_VID_SId_frameNo*. Here, the first four components are the same as described above, while the last component, $frameNo$, corresponds to the

TABLE IV: Segment-wise rates of Viola-Jones, Fast Face, Tiny Face and Single Shot Scale-Invariant Face detectors.

Spectrum	Algorithm	Distance A (Farthest)		Distance B (Intermediate)		Distance C (Closest)		Overall	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
VIS	**Viola Jones [24]**	0.43	0.91	0.42	0.97	0.37	0.96	0.40	0.94
	Fast Face [16]	0.22	0.84	0.22	0.96	0.19	0.97	0.21	0.92
	Tiny Face* [10]	0.98	0.97	0.96	0.99	0.93	0.99	0.96	0.99
	S^3FD [31]**	0.95	0.70	0.95	0.96	0.36	0.95	0.60	0.87
NIR	**Viola Jones [24]**	0.79	0.58	0.80	0.79	0.79	0.90	0.79	0.76
	Fast Face [16]	0.19	0.10	0.35	0.41	0.44	0.70	0.37	0.40
	Tiny Face* [10]	0.99	0.95	0.97	0.99	0.95	0.99	0.97	0.98
	S^3FD [31]**	0.92	0.52	0.95	0.86	0.44	0.96	0.65	0.78

*at best thresholds of 0.18 and 0.21 in VIS and NIR spectra respectively. **at best thresholds of 0.29 and 0.25 in VIS and NIR spectra respectively.

frame number to which the annotation belongs. For example, *N_S4_V87_126_47* refers to the face annotation of subject ID 126, for the video N_S4_V87, at frame number 47.

C. Division into Training and Testing

A fixed training and testing protocol has been provided with the proposed dataset. Videos corresponding to 200 subjects form the testing set, while data pertaining to the remaining 52 subjects form the training partition. This results in a training and testing set of 72 and 388 videos, respectively. Out of the 200 test subjects, 8 belong to the open-set, that is, their gallery images are not made available. In order to simulate a real world scenario, it is ensured that the training and testing partitions are subject-disjoint and video-disjoint.

III. Face Detection

Face detection refers to the task of finding faces in a given image. It is a crucial step for many face analysis applications such as face alignment, face tracking, face recognition, and face verification [28]. A highly effective and popular technique for face detection was presented by Viola and Jones [24], wherein Haar-like features are extracted and adaptive boosting is used to train a cascade of classifiers. More recently, research in this domain is based on deep learning algorithms [10], [21], [31], presented in detail by Zafeiriou *et al.* [30]. In order to perform baselining on the proposed dataset, we compute face detection results using some popular as well as state-of-art face detectors, namely, Viola Jones Face Detector [24], Fast Face Detector[16], Tiny Face Detector [10], and Single Shot Scale-Invariant Face Detector (S^3FD) [31].

A. Face Detection Protocol

In order to analyze the effect of distance for face detection, results are reported on the entire video ($1-36ft$) as well as distance wise. The videos are divided into three parts based on the distance of the subject from the camera, namely distance A (farthest, $24-36ft$), distance B (intermediate, $12-24ft$) and distance C (closest to the camera, $0-12ft$). This is done for videos belonging to both spectra - visible (VIS) and Near Infrared (NIR). Results are computed by comparing the bounding boxes returned by the face detectors to the annotated ground truth detected faces. A detected face is considered a true positive if the bounding box and the annotated face are overlapping, else it is counted as a false positive. Experimental evaluation has been done on the test set (described in Section II-C) and baseline results corresponding to the four face detectors have been reported on the same.

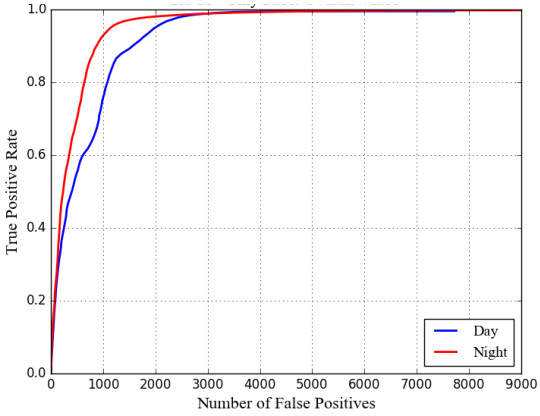


Fig. 4: ROCs for Tiny Face - the detector giving best detection results on proposed dataset, computed over entire video in VIS (Day) and NIR (Night) spectra respectively.

B. Results

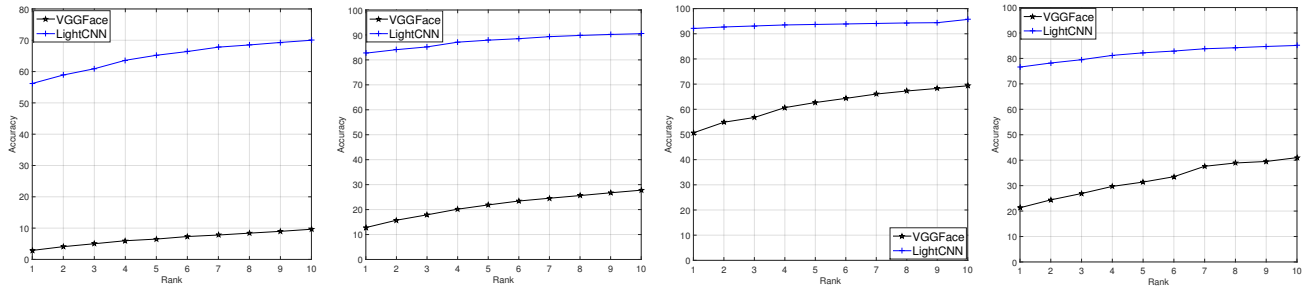
Table IV presents the face detection precision and recall rates computed over the entire video and the three video segments for NIR and VIS scenarios using four face detectors. The best results are reported using the Tiny Face detector, which is pre-trained on low resolution faces. However, it is important to observe that apart from Tiny Face, other face detectors achieve at most 0.76 precision for the entire video, thereby depicting the challenging nature of the database. The low face detection performance can be attributed to the unconstrained settings resulting in low resolution, occlusion of faces, and non-uniform illumination. Fig. 4 presents the Receiver Operative Characteristics (ROC) curves for both VIS (day) and NIR (night) videos using the Tiny Face detector. It can be observed that even the best performing face detector obtains a high number of false positives which demonstrates the challenging nature of the database. According to the ROC curves, Tiny Face detector performs better in the NIR spectrum (night-time), as opposed to the visible spectrum (day-time). This can be attributed to the fact that the night-time videos have high illumination on the subject, with limited background information, whereas the day-time videos contain a large amount of well illuminated background, resulting in higher false positives.

TABLE V: Rank-1 accuracies (%) for identification obtained by running COTS-A and COTS-B for the two experimental protocols: a) Distance Wise in columns 3 to 5, and b) Entire videos as probe in column 6.

System	Experiment	Distance A	Distance B	Distance C	Overall Video
COTS-A	**CR**	8.18	37.18	72.12	39.63
	CS-CR	0.09	1.56	11.23	4.38
Verilook (COTS-B) [1]	**CR**	9.44	38.07	74.84	41.26
	CS-CR	0.69	2.783	25.61	9.90

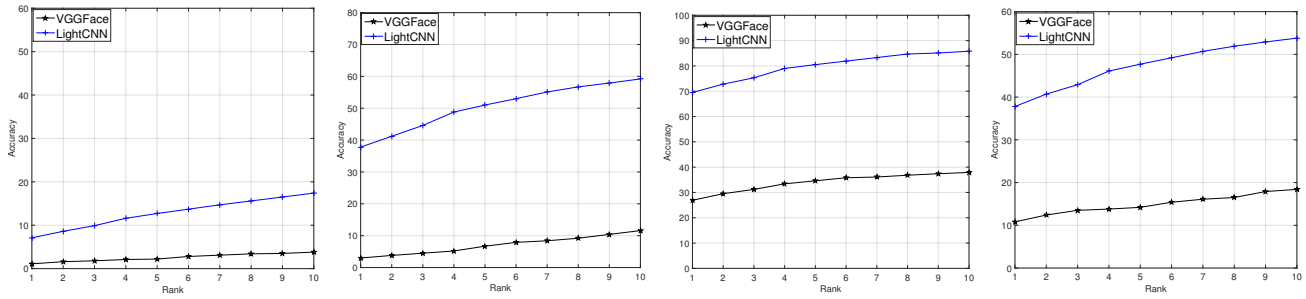
TABLE VI: Rank-1 accuracies (%) for identification base-lining results obtained by running deep learning models for the two experimental protocols: a) Resolution Wise in columns 3 to 5, and b) Entire videos as probe in column 6.

Algorithm	Experiment	Distance A	Distance B	Distance C	Overall Video
LightCNN29 [27]	**CR**	56.24	82.76	92.16	76.04
	CS-CR	7.19	37.85	69.54	37.86
VGGFace [20]	**CR**	2.87	12.78	50.41	21.35
	CS-CR	1.18	3.09	26.98	10.87



(a) Distance A (b) Distance B (c) Distance C (d) Overall video

Fig. 5: CMC curves for visible spectrum (day-time) videos: CR face recognition for different resolutions and overall videos.



(a) Distance A (b) Distance B (c) Distance C (d) Overall video

Fig. 6: CMC curves for NIR spectrum (night-time): CS-CR face recognition for different resolutions and overall videos.

IV. Face Recognition

Face recognition in controlled scenarios is a comparatively easier problem, while the same in unconstrained settings continues to be a major research problem [17], [23]. Modern surveillance systems are composed of two distinct sections, namely the surveillance cameras which are used to capture the probe/query images, and a background database (or watch list) known as the gallery. The proposed database is composed of daytime (VIS) and night-time (NIR) probe videos, and a separate gallery captured in visible spectrum under controlled scenarios. Each video contains the subject at a standoff distance of $0-36ft$, thus in those frames where the subject is at a larger standoff distance, the effective resolution of the detected faces would be lower. The proposed FaceSurv database is evaluated on two heterogeneous face recognition problems, namely cross-resolution (CR) face recognition and cross-spectral cross resolution (CS-CR) face recognition. Additional challenges such as variation in pose and illumination, occlusions, image quality and expressions make the problem much more challenging. The proposed database simulates both cross-resolution and cross-spectral characteristics of actual surveillance scenarios. We present results of baseline experiments using two commercial matchers namely COTS-A and Verilook [1] (COTS-B), and two state-of-the-art deep convolutional networks namely LightCNN29 [27] and VGGFace [20] on both the face recognition scenarios (CR and CS-CR).

A. Face Recognition Protocol

We perform two experiments - Cross-Resolution (CR) where frames of day videos (visible spectrum) are matched with the gallery and Cross-Spectral Cross-Resolution (CS-CR) where frames of night videos (NIR) are matched with

the gallery. The images captured in controlled scenarios (in high resolution and visible spectrum) corresponding to each of the subjects which do not belong in the openset serve as our gallery images. This results in 576 gallery images, corresponding to 192 subjects (3 each). Subjects in the openset do not have gallery images available. For experiments with the commercial matchers, the annoatetd face images of the FaceSurv dataset are used as probe images, on the other hand, for VGGFace and LightCNN29, Viola-Jones [24] face detector was used on the annotated frames for detection of faces for all the probe videos.

*1) **Experiment I: Cross-Resolution (CR)**:* In this experiment, the detected faces from the day video frames are considered as probes. Each video is divided into 3 parts, based on the standoff distance of the subject from the camera (as explained in section III-A). In order to report results for the deep CNN models, detected face images pertaining to distance A are resized to 24×24, those for distance B are resized to 48×48 and those for distance C are resized to 96×96. The farther the standoff distance, the lower is the effective resolution of the face. Thus, the experiment of matching low resolution visible spectrum probes with high resolution visible spectrum gallery is a cross resolution face recognition problem. The results are reported as follows:

1) **Distance-Wise:** Identification is performed across all video frames, where each frame belongs to one of the 3 parts of the video depending on the standoff distance of the subject from the camera (as explained in section III-A)
2) **Entire Video:** Identification is performed for all the video frames individually.

*2) **Experiment II: Cross-Spectral Cross-Resolution (CS-CR)**:* In this experiment, the detected faces from the night video frames are considered as probes. There are 60,644 annotated faces in total from 189 videos. Similar to experiment I, we have 576 still images which serve as our gallery corresponding to 192 subjects (3 each), among the total 200 test subjects. These probes are in the NIR whereas the gallery images are in the visible spectrum. In addition to that, the factor of distance from the camera makes this a cross-spectral cross-resolution face recognition experiment.

While baseline results have been reported for the above mentioned protocols, it is important to note that the proposed FaceSurv dataset provides the flexibility of performing matching with other protocols as well. For example, keeping the training and testing partitions consistent, results can also be reported video to video matching, both within spectrum (VIS-VIS) and between spectrum (VIS-NIR).

B. Results

Table VI shows the Rank-1 identification accuracies (%) for both CS-CR and CR. We use two commercial matchers namely COTS-A and Verilook (COTS-B). Out of 576 gallery images, 574 images are enrolled successfully for COTS-A, whereas all the gallery are enrolled successfully for COTS-B. Key highlights of the results are as follows:

Fig. 7: Failure cases for face recognition, *top row:* NIR spectrum (night-time), *bottom row:* visible spectrum (daytime) videos. Low resolution and poor illumination results in challenging face images.

(i) Identification Performance: Table VI presents the rank-1 identification accuracy with different feature extractors and COTS. LightCNN29 achieves the best performance across all comparative techniques by reporting an accuracy of 76.05% and 37.86% for the scenario of CR and CS-CR, respectively. Fig. 6 and 5 presents the Cumulative Match Characteristic (CMC) curves for both the scenarios. It is interesting to observe that while LightCNN29 is able to achieve around 85% identification accuracy at rank-10 for cross resolution scenario, it achieves around 52% rank-10 accuracy for cross spectral cross resolution recognition. This demonstrates the challenging nature of the proposed dataset, and strengthens the requirement for dedicated research in this direction.
(ii) CR vs CS-CR: Since CS-CR is a more challenging problem than CR face recognition, the identification accuracies for CR face recognition are better than that of CS-CR face recognition across all the recognition algorithms. In addition to that, a large the standoff distance of the subject from the camera reduces the effective resolution of the detected face image. It can be observed that across all four algorithms, the recognition performance for distance C is cumulatively higher than in distance B followed by distance A, as those corresponds to much lower probe resolutions. Even LightCNN [27], which achieves excellent results on CR face recognition (92.16% in distance C) performs poorly for CS-CR (only 7.19% in distance A). Fig. 7 presents sample face images incorrectly recognized by most of the techniques. The images demonstrate the challenging nature of the proposed dataset, wherein variations due to distance, spectrum, illumination, motion blur and occlusion render face recognition challenging.
(iii) COTS vs Deep CNNs: LightCNN [27] is the best performing model across all the four algorithms for both CR and CS-CR face recognition. For overall video, LightCNN [27] and VGGFace perform better for CS-CR face recognition as compared to the COTS. However, for CR face recognition VGGFace performs worse than both the COTS methods.
(iv) Performance across different distances: It is noteworthy that even at a small standoff distance from the camera (for distance C) for CR face recognition, the results are not high enough. In addition to that, when the subject is at distance A

or distance B, the results are poor, owing to low resolution and quality, poor and uncontrolled illumination and high domain difference with respect to the gallery images. It can be seen that across all the four algorithms, the performance for distance A is lower (especially in night-time videos) than the other two distances.

V. Conclusion

Research in face recognition for surveillance scenarios remains a long standing problem, further aggravated by the availability of limited datasets. In this research a novel dataset, FaceSurv is proposed for facilitating research in this direction. The proposed dataset contains 460 videos pertaining to 252 subjects, captured across multiple sessions and spectra. Ground truth annotated face regions and high resolution gallery images are also available, along with a fixed protocol for performing face detection and face recognition. Baseline results have been reported for the task of face detection with four algorithms: Viola Jones, Fast Face Detector, Tiny Face Detector, and Single Shot Scale-Invariant Face Detector. Similarly, baseline face recognition results have been reported with two Commercial Of The Shelf systems, namely COTS-A and Verilook (COTS-B), and two state-of-the-art deep convolutional networks namely LightCNN29 and VGGFace. Results across face detection and recognition demonstrate the complex and challenging nature of the proposed dataset. In order to facilitate research in this direction, the proposed dataset will be released for research purposes. We believe that the research community will benefit from the proposed FaceSurv dataset for developing robust face recognition systems applicable in real world surveillance scenarios.

VI. Acknowledgement

We thank the volunteers for taking part in the data collection. This research is partially supported through MEITY (Government of India), India, and the Infosys Center for Artificial Intelligence, IIIT-Delhi. S. Ghosh and S. Nagpal are supported via the TCS PhD fellowship.

References

[1] Verilook. {http://www.neurotechnology.com/verilook.html}.

[2] 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.

[3] B. R. Beveridge, P. 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, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *Biometrics: Theory Applications and Systems*, pages 1–8, 2013.

[4] 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.

[5] 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.

[6] 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.

[7] T. I. Dhamecha, M. Shah, P. Verma, M. Vatsa, and R. Singh. CrowdFaceDB: Database and benchmarking for face verification in crowd. *Pattern Recognition Letters*, 107:17 – 24, 2018.

[8] R. Goh, L. Liu, X. Liu, and T. Chen. The CMU face in action (FIA) database. In *International Conference on Analysis and Modelling of Faces and Gestures*, pages 255–263. 2005.

[9] M. Grgic, K. Delac, and S. Grgic. SCface – Surveillance Cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, 2009.

[10] P. Hu and D. Ramanan. Finding tiny faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[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] N. D. Kalka, B. Maze, J. A. Duncan, K. A. OConnor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. IJB–S: IARPA Janus Surveillance Video Benchmark. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2018.

[14] 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.

[15] 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.

[16] S. Liao, A. K. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):211–223, 2016.

[17] Z. Lu, X. Jiang, and A. C. Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 2018.

[18] 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.

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

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[21] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[22] M. Singh, S. Nagpal, N. Gupta, S. Gupta, S. Ghosh, R. Singh, and M. Vatsa. Cross-spectral cross-resolution video database for face recognition. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2016.

[23] M. Singh, S. Nagpal, M. Vatsa, R. Singh, and A. Majumdar. Identity aware synthesis for cross resolution face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 479–488, 2018.

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

[25] 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.

[26] 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.

[27] X. Wu, R. He, and Z. Sun. A lightened CNN for deep face representation. *CoRR*, abs/1511.02683, 2015.

[28] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[29] 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.

[30] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, 138:1 – 24, 2015.

[31] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *IEEE International Conference on Computer Vision*, 2017.

[32] 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.