IJB–S: IARPA Janus Surveillance Video Benchmark *

Nathan D. Kalka † Brianna Maze † James A. Duncan † Kevin O'Connor ‡
Stephen Elliott ‡ Kaleb Hebert † Julia Bryan ‡ Anil K. Jain §

Abstract

We present IJB–S dataset, an open-source IARPA Janus Surveillance Video Benchmark and associated protocols. The dataset consists of images and surveillance video collected from 202 *subjects at a Department of Defense (DoD) training facility. Surveillance video was captured across multiple vignettes representative of a variety of real-world surveillance use cases that are particularly of interest to law enforcement and national security communities. Each video was annotated by human subject matter experts in order to generate ground truth identity and bounding box face labels. In total, over 10 million annotations were collected for the dataset. We present benchmark results utilizing state of the art deep learning approaches such as FaceNet. Our results illustrate and characterize the difficulty of the dataset.*

1. Introduction

Research on face recognition from video has increased in intensity in recent years due to law enforcement and commercial applications, advances in camera sensing technology, and improvement in face recognition technology due to the proliferation of deep learning algorithms. The unique properties inherent to videos enable both humans and automated systems to accurately perform recognition in challenging viewing conditions. Although interest in video based face recognition has increased, significant research challenges remain as video based media is typically of inferior quality due to the lack of constraints present in the capture environment. Research on video based face recognition has also been hindered by the lack of publicly available datasets that are representative of real-world environments.

In 2014, there was an estimated 245 million[1] surveillance cameras active and operational globally. Due to diverse and important use cases including identifying terrorists, home security, and rescuing victims of child exploitation, face recognition from video has garnered much attention in the computer vision and biometric research communities. Video surveillance systems are designed to be unobtrusive, so the activities of the recorded individuals and the effects of the environment can vary significantly. This coupled with low resolution capture by deployed surveillance cameras, leads to poor quality face images in terms of pose, illumination, and expression (PIE) variations.

In traditional face image acquisition environments, such as passport agencies or law enforcement booking stations, nuisance variables including head pose and facial expression, are constrained during collection. While state of the art face recognition systems are performing near-human levels of accuracy on constrained face imagery, their performance generally falls short in comparison to human performance on unconstrained media [2, 4, 23].

This performance gap for video surveillance imagery has motivated the development of face recognition algorithms that compensate for the difficult conditions encountered in uncontrolled viewing scenarios. Indeed, videos inherently provide additional information [1] that can be exploited to aid recognition tasks. Unlike still imagery, videos contain temporal information that can be utilized for improved recognition performance. Furthermore, a sequence of video frames can display the same object from a number of different viewing angles, thus 2D videos can be used to construct 3D subject-specific face models. The combination of various information sources that video data provides can be exploited to build a robust face representation in unconstrained operating environments.

In order to continue development and advance uncon-

*This research is based upon work supported by the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA), via FBI Contract #DJF-16-1200-G-0009392. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

†N. D. Kalka, B. Maze, J. A. Duncan, and K. Hebert are with Noblis, Bridgeport, WV, U.S.A.

‡K. O'Connor, S. Elliott, and J. Bryan are with Purdue University, West Lafayette, IN, U.S.A

§A. K. Jain is with Michigan State University, East Lansing, MI, U.S.A.

978-1-5386-7180-118$31.00

[1] https://technology.ihs.com/532501/245-million-video-surveillance-cameras-installed-globally-in-2014

Dataset	# subjects	avg. # vid/subj	# annotations	avg. fps	resolution
IJB-S	202	12	>10M	30	1280×960 to $2592 \times 1520^{**}$
ChokePoint[25]	25	2	64,204	30	800×600
PaSC[3]	265	11	248,637	–	640×480 to 1280×720
COX [11]	1,000	3	–	–	66×66 to 798×798
Youtube Celebrities [12]	47	41	1,910	25	180×240 to 240×320
Honda/UCSD [14]	20	3	–	15	640×480
MBGC [19]	821	5	–	30	1440×1080
UCCS [8]	1,732	N/A†	>70,000	1^{*}	5184×3456
IJB-C[15]	3,531	3	>3.3M	30	61×77 to 14400×9600
LFW[10]	5,749	N/A†	13,233	N/A	250×250
MF2[16]	672,057	N/A†	4.7M	N/A	72×72 to 29905×6432
MS-Celeb-1M [9]	100,000	N/A†	10M	N/A	–
SCface [7]	130	N/A†	4,160	N/A	426×320 to 680×556
Quis-Campi [17]	320	6	–	–	–
Stallkamp et al. [22]	41	56	–	–	–

Table 1. A comparison of IJB-S to other unconstrained video based datasets. Entries in the table with "–" indicate that information was not provided. *UCCS only releases images taken from a camera @ 1 fps. **IJB-S also includes still images captured from a fixed wing UAV at a resolution of 4552×3292. † These datasets do not include videos.

strained face recognition technology, researchers must have access to large amounts of relevant training and testing data with reliable ground truth information. This data must also be accompanied with operationally relevant testing protocols to facilitate repeatable and reproducible assessments. The objective of this paper is to provide such a dataset, IJB-S, and the testing protocols with baseline performance to the research community[2].

1.1. Background

Table 1 provides a summary of unconstrained video based datasets available in the public domain. Example imagery from selected datasets can be seen in Fig. 1. One of the most popular datasets for evaluating face recognition technology in surveillance scenarios is the ChokePoint dataset [25]. ChokePoint was designed to test performance in 1:1 verification scenarios. Media was captured using three surveillance cameras placed above natural chokepoints to mimic surveillance scenarios. The dataset is limited by its number of subjects, number of video sequences, and its lack of variation in capture environment (all video was captured indoors).

In 2013, NIST released the Point and Shoot dataset (PaSC) [3] to facilitate research in unconstrained environments. While this dataset includes stills and videos in both indoor and outdoor environments, the dataset was curated utilizing hand-held camcorders, which provide systematically different fields of view and photometric properties than those of surveillance cameras. The dataset also only provides 1:1 verification protocols.

[2]This database, like previous IJB databases, will be released in the public domain in summer 2018

Table 1 provides other datasets that have been released to advance research in the area of surveillance video face recognition, but they are also limited by a variety of factors. The COX dataset [11] includes 3,000 video sequences from 1,000 subjects, and utilizes hand-held camcorders to capture videos in an indoor environment. Youtube Celebrities [12] includes 1,910 video sequences of 47 subjects, manually curated from Youtube. Although COX [11] and Youtube Celebrities [12] contain a wide variety of pose, illumination, and expression, the media is more photo-journalistic in nature and not representative of real-world surveillance scenarios and environments for law enforcement and national security communities. Lee et al. [14], released a dataset of 20 subjects and 52 video sequences, but the small number of subjects limits its usage for training and evaluation of algorithms. The dataset in [19] is much larger, but the group scenarios and subject interactions are limited. The UCCS dataset [8] consists of still images only. The remaining datasets, IJB-C[15], LFW[10], MF2[16], and MS-Celeb-1M[9] are datasets that include media scraped from the web, and as such, contain celebrities and scenarios not representative of surveillance scenarios.

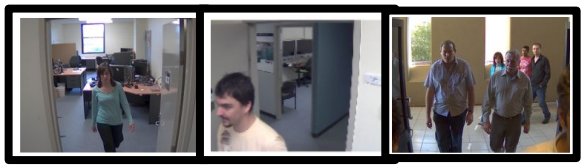
While the above datasets have been instrumental in advancing the state of the art, they are limited in terms of real-world capture environments, variety of subject activities and interactions, and annotations and protocols that can accurately measure end-to-end performance (e.g. joint detection, clustering, and recognition) of face recognition systems. To remedy the deficiency, this paper introduces the IARPA Janus Benchmark–Surveillance (IJB-S) dataset, which contains a corpus of annotated, unconstrained face videos and surveillance relevant protocols to push the frontiers of unconstrained face recognition.

2. IJB-S Dataset

The IARPA Janus Benchmark – Surveillance (IJB-S) dataset contains still images and videos for 202 subject identities. The data was collected across two weekends in November 2017 at a Department of Defense (DoD) training facility. Forty participants repeated the collection across both weekends. This venue was chosen because it offers many training structures that can be used to simulate law enforcement and national security use cases for video surveillance collection.

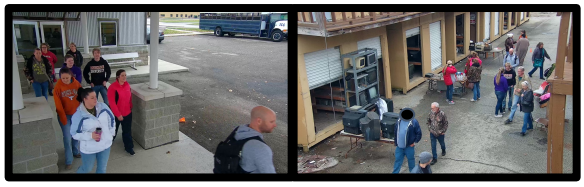
IJB-S includes 350 surveillance videos spanning 30 hours in total, 5,656 enrollment images, and 202 enrollment videos. On average, each subject is present in 12 surveillance videos. All visible faces in each image or frame have corresponding bounding box coordinates annotated by human experts. In total, over 10 million manual annotations were produced during the curation of the IJB-S data set.

(a) Point and Shoot frames



(b) ChokePoint frames

(c) IJB-S Panasonic WV-SW395 frames



(d) IJB-S Speco O4P30X frames

Figure 1. An illustration of imagery and frames from popular video datasets in the academic literature: PaSC [3], ChokePoint [25], and IARPA Janus Benchmark. Sample fields of view (FOV) from all of our collection vignettes can be found in Fig. 2(b)

The contributions of the IJB-S data set to the face recognition and biometrics communities are the following:

- Subjects with full pose variation, including extreme pitch and yaw;
- Videos with a variety of resolutions, motion artifacts, and standoff distances to represent the spectrum of quality of real-world surveillance cameras;
- Image- and frame-specific bounding box annotations;
- Protocols for face detection, 1:N identification (supporting open- and closed-set evaluation), and end-to-end face recognition system evaluation;
- Benchmark accuracy measures from a Government-Off-The-Shelf (GOTS) algorithm and state-of-the-art face recognition algorithms that utilize deep neural networks (FaceNet [20]);
- Clear authority for redistribution via approved Institutional Review Board collection plan.

2.1. Collection Methodology Overview

The DoD collection campus includes a wide range of training structures for simulating real-world use cases, seen in Fig. 2. The School on the campus served as the staging area for participant enrollment and the base of operations for data collection administrators. The subway station, bus station, embassy, and marketplace areas were selected as vignettes[3] due to their ability to represent operational environments and further enable the study of crowd dynamics in video and re-identification scenarios. Additional locations on the campus were selected for placement of surveillance cameras that recorded the navigation path of participants as they progressed through the video collection process.

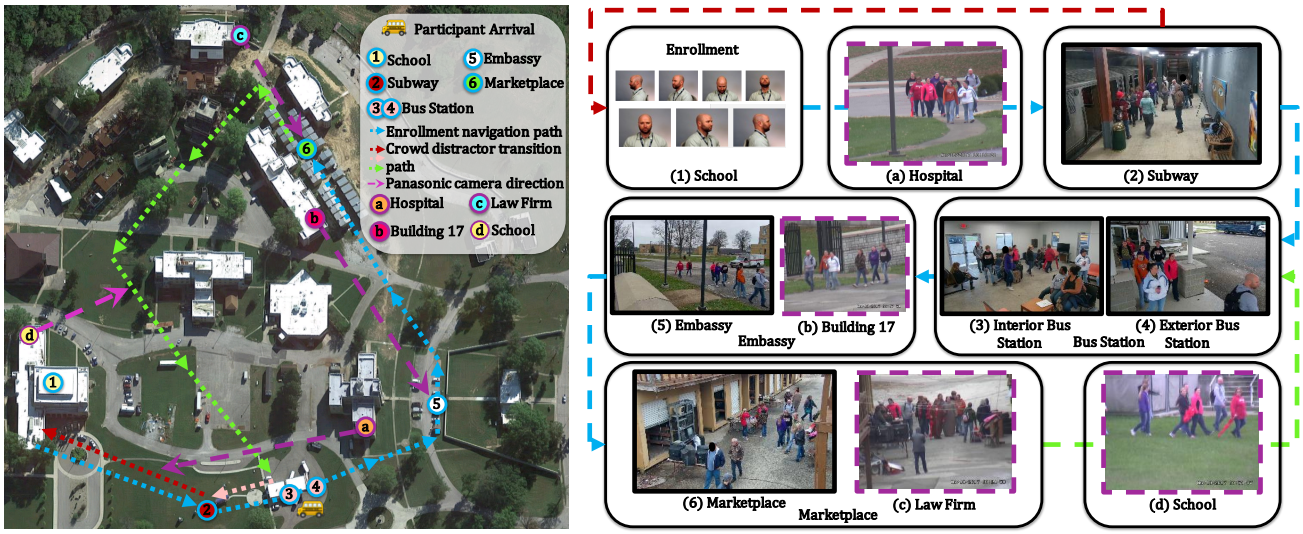
Participants were consented off-site which included reading, reviewing, and signing an Institutional Review Board (IRB)-approved consent form. After consent, participants[4] arrived in groups of 40 at the bus station. After a short explanation on the collection process, instructions were provided and participants were then split into four groups of 10: an "enrollment group", and three crowd distractor groups. Each of the three distractor groups headed to the subway station, bus station interior, and outdoor marketplace, and served as crowd distractors for the enrollment group as they progressed through the collection. Collection administrators were present with the groups throughout the duration of the collection.

2.1.1 Enrollment

The enrollment group was directed towards the School for enrollment. Here, participants completed a form solicit-

[3] Vignettes represent collection scenarios involving structures or props that would be of interest to real-world surveillance applications.

[4] For this collection, subjects were recruited through a local staffing agency. The subjects were paid $50/hr for their participation in this collection.



(a) Bird's Eye View of the Collection Protocol

(b) Walk-Through of the Collection Protocol with Associated FOVs

Figure 2. Participant walking paths for both the enrollment and crowd distractor groups are illustrated in Fig. 2(a). The rooftop surveillance camera directions are displayed as purple dashed arrows. The FOVs for each mounted camera can be seen in 2(b). Prior to entering the marketplace, participants had the opportunity to wear prop clothing, increasing session variability.

ing demographic information such as age, race, and gender. After completing the form, still images and video were collected from the subjects in a constrained laboratory setting. In particular, the setup loosely followed the subject acquisition protocol 50 (SAP50) requirements defined in ANSI/NIST-ITL 1-2007 [18]. This included the use of 3-point lighting, an 18% gray backdrop, and a standoff distance of 2 meters. Still images were captured from both a Sony DSC-H300 point-and-shoot and a Samsung Galaxy 5 cellular phone. A Canon Vixia HF R400 handheld camcorder was also utilized to collect 1080P video of the participants at they navigated through each shot.

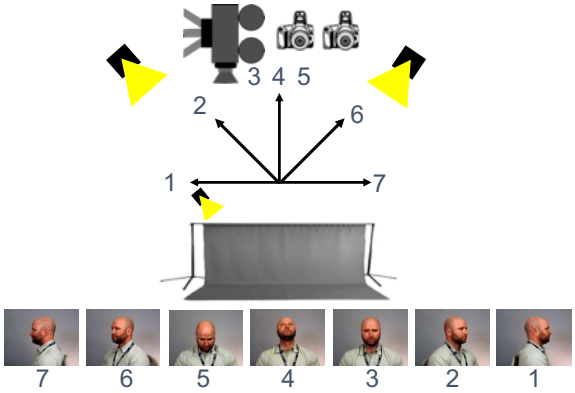


Figure 3. An illustration of the enrollment equipment setup for the collection. Still imagery was captured from both a Sony DSC-H300, and a Samsung Galaxy 5 consumer cell phone. Video enrollment data was collected with a Canon Vixia HF R400 camcorder.

In addition to frontal, $\pm 45°$, and $\pm 90°$ poses, images were collected with $\pm 20°$ pitch angle. To assist the participants and reduce collection overhead time, high visibility tape was placed on the floor at the different yaw angles and the ceiling for pitch. Participants sat on a rotating stool allowing efficient and repeatable iteration through the different shots. In total, enrollment media collection took approximately 2 minutes per subject. An illustration of this process is provided in Fig. 3. Overall, 37% of the participants were male. The average age across all participants was 38.8. Each participant has an average of 18K bounding box annotations.

2.1.2 Collection Protocol

Once enrollment was complete, the enrollment group was led by collection administrators on a pre-determined route that visited all four vignettes, the path of which can be seen in Fig. 2(a). In short, the enrollment group transitioned through the subway, interior and exterior bus station, embassy, and completed the route at the marketplace.

The enrollment group spent a short period of time at each vignette, mingling with the crowd distractors posted there. Each vignette was under surveillance by at least one mounted camera. Additionally, the same subjects appeared in multiple videos at different viewpoints, standoff distances, etc., due to routes crossing the fields of views of multiple surveillance cameras. The FOVs for each camera are outlined in Fig. 2(b).

Once the enrollment group completed their transition to the marketplace, the participant groups rotated their roles and locations as follows:

- Enrollment group remained at the marketplace and became the marketplace crowd distractor group;
- Marketplace crowd distractor group became bus station crowd distractor group;
- Bus station crowd distractor group became subway crowd distractor group;
- Subway crowd distractor group transitioned to the School and became the enrollment group.

This cycle continued until all groups had served as the enrollment group. It took approximately 3 hours for a group of 40 participants to rotate through all the locations. At that point, the collection ended and all participants boarded the bus to be transported off the campus.

2.2. Surveillance Cameras

Commercial off-the-shelf surveillance cameras were utilized for the video collection, specifically Panasonic WV-SW395[5] and Speco O4P30X[6] dome cameras. During the last weekend of the collection, a small fixed-wing unmanned aerial vehicle (UAV) was flown over the collection area, specifically the marketplace, opportunistically capturing surveillance video. Fig. 5 displays extracted video frames and a high resolution still image captured from the UAV.

The Panasonic dome PTZ surveillance cameras were provided and setup to specifications by base personnel. The collection administrators designed custom mounting stands for the Speco cameras as they could not be affixed to any of the standing structures on the campus. Inclement weather was another consideration when designing the custom mounts, and the collection team weatherproofed all electronics since several of the cameras would be placed outside and needed to withstand strong gusts of wind and rain without significant impact to video quality. An illustration of the mounting stands and weatherproofed electronics is provided in Fig. 4.

At the beginning of each collection day, camera settings such as field of view, frame rate, and card space were calibrated for consistency and to ensure smooth operation. Table 2 lists specifications and settings utilized for the surveillance cameras. Each camera recorded to a 64GB or 128GB compact flash memory card. Since there were only two collection sessions per day, the surveillance cameras were set to record before the morning session and recorded continuously throughout the day until the afternoon session completed. At the end of the afternoon session, the cameras were turned off and the flash memory cards were retrieved in order to create a backup of the data for the day.

The standoff distances of the cameras varied depending on the vignette and whether the camera was indoor or outdoor. Overall, the minimum standoff distance was approximately 5ft at the bus and subway stations while the longest was at the embassy, via a Panasonic mounted on the roof of Building 17 whose field of view contained the embassy gate. Fig. 2(b) provides an illustration of the fields of view for each of the surveillance cameras.

[5]https://security.panasonic.com/products/wv-sw395/

[6]http://specotech.com/index.php/products/video/cameras/ip/item/1251-o4p30x

(a) Mount (b) Electronics

Figure 4. Custom mounting stands and weatherproofing for the Speco camera and electronics. The stands can be raised up to a height of 10ft. We attached the head piece to buckets filled with 60lbs of sand via 1000lb steel cable to mitigate strong winds.

Camera	Model	Resolution	Frame Rate	Location
1	Panasonic PTZ WV-SW395	720p	30fps	outdoor
2	Speco PTZ O4P30X	2592×1520	30fps	both
3	UAV	720p	30fps	outdoor

Table 2. A summary of surveillance camera specifications. Four Panasonic cameras were placed on the rooftops of the school, hospital, law firm, and Building 17. Six Speco cameras were setup in the subway, bus station, embassy, and marketplace areas. The UAV collected both 720p video and 4552×3292 stills.

2.3. Post-Collection Annotation

Prior to annotating bounding boxes and the ground truth labeling of subjects, the surveillance videos were split into clips based on when activity was present in the field of view. This was done automatically for the the videos captured by the Speco cameras utilizing a tool called SuBSENSE [21] to identify temporal regions of interest. Videos were then split based on the identified regions of interest and manually reviewed for correctness. We could not find a good set of SuBSENSE parameters for videos captured by the Panasonic cameras, thus temporal regions of interest for those videos were manually identified.

The enrollment and surveillance video media were then annotated with ground truth subject identity labels and facial bounding boxes. Due to the nature and difficulty associated with the data, the entire annotation process was performed in-house as opposed to on a crowd source platform such as Amazon Mechanical Turk. Using a small pool of annotators knowledgable in the domain ensured consistent quality across the annotations. In addition, the collection

Figure 5. Sample frames extracted from video collected by a small fixed-wing UAV. The UAV circled the marketplace area and captured surveillance video. The center image is a high resolution still.

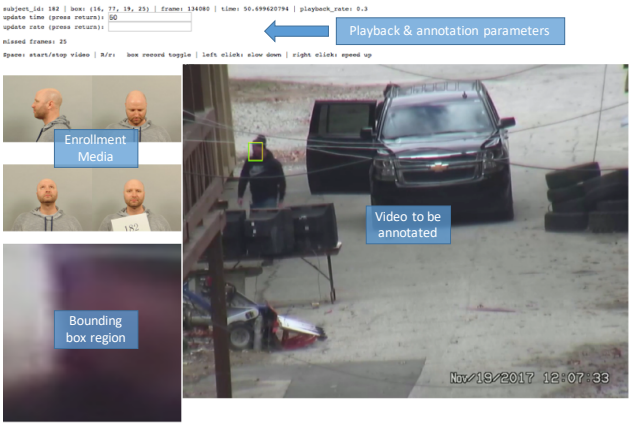


Figure 6. A screenshot of the user interface for annotation. Domain experts track the subject through the video annotating facial bounding boxes. License plates and faces of individuals who have not been consented are blurred.

administrators became familiar with the subject participants by the second day of data collection and could readily identify them. Annotation was facilitated through a web-based tool with a graphical user interface developed by the collection administrators. A screenshot of the annotation interface can be found in Fig. 6.

In total, over 10 million manual annotations were collected in the curation of IJB-S over a six week period. This annotation process produced (i) an accurate ground truth corpus of videos containing bounding boxes for face detection evaluations and (ii) subject labels for face recognition.

3. Evaluation Protocols

Table 3 outlines key statistics of the protocols to be released with IJB-S. For the 1:N identification protocols there are two disjoint galleries, G1 and G2 with 101 subjects each. Each subject is enrolled into a single template within one of the galleries. The galleries are disjoint from each other so that open-set identification scenarios (i.e., searches where the probe template does not have a corresponding mate in the gallery) can be tested. All query templates are searched against both G1 and G2.

IJB-S includes a face detection protocol and several open-set 1:N identification experiments: surveillance-to-still, surveillance-to-booking, multi-view surveillance-to-booking, surveillance-to-surveillance, and UAV surveillance-to-booking, where the "booking" template comprises the full set of images taken of a subject at enrollment time. Throughout these protocols, IJB-S utilizes the concept of subject-specific modeling [13, 24, 15], in which a single template is generated for a subject based upon the available pieces of media, a paradigm shift from the traditional process of creating a template for every available piece of media (i.e. still image or frame).

Experiment	Test Name	Reference Media	Query Media	# Bounding Boxes	# Images	# Videos
1	Face Detection in Video	N/A	N/A	3,665,339	N/A	N/A
2	Surveillance-to-still	Single Frontal Still	Surveillance Video	N/A	202	179
3	Surveillance-to-booking	SAP50 Still Images	Surveillance Video	N/A	1,414	179
4	Multi-view Surveillance-to-booking	SAP50 Still images	Multiple Surveillance Videos	N/A	1,414	179
5	Surveillance-to-Surveillance	Surveillance Video	Surveillance Video	N/A	N/A	381
6	UAV Surveillance-to-booking	SAP50 Still Images	UAV Videos	N/A	1,487[7]	10

Table 3. Testing protocols for IJB-S: face detection in video and multiple open-set 1:N experiments involving surveillance video. The "booking" reference template comprises the full set of images taken of a subject at enrollment time.

3.1. Performance Metrics

The identification protocols are end-to-end protocols which require joint face detection, clustering, and recognition. These protocols are evaluated according to the two metrics described below: (i) End-to-End Retrieval Rate (EERR) and (ii) a variant on the Identification Error Trade-off (IET) [24] as in [15].

The EERR evaluates accuracy in a closed-set identification scenario relative to rank. The EERR expresses the proportion of mated searches returning a match at or above a particular rank, where a mated search is defined as having a corresponding mate in the gallery. Note that there are two scenarios in which a mated search may result in a miss: (1) the face of the subject of interest was not detected, or (2) the face of the subject of interest was detected but the resulting candidate list did not contain the mate. A further modification from the standard Cumulative Match Characteristic curve is the use of subject weighting on a media-by-media basis. Specifically, subjects in videos can receive a hit or miss of ± 1. Thus, detections in frames are weighted such that all of a subject's frames from a single video have a weight of one. For instance, if a detector correctly detected only 25 of 100 subject frames, that resulted in a hit at rank, r, then the weight would be hit(r) = [hit(r) + 25/100]. Weighting in this manner can be used to further differentiate algorithms based on their ability to generate accurate subject video tracks.

A correct detection is defined as a (normalized) predicted bounding box which has an intersection over union score of at least 50% with the ground truth bounding box. Predicted boxes are rigidly normalized by increasing or decreasing the area of the predicted box until it matches the area of the ground truth bounding box. If the area of the predicted box differs from that of the ground truth box by over 150%, then the predicted box is automatically considered to be a false alarm.

The IET expresses how the False Positive Identification count (FPI) varies with respect to the False Negative Identification Rate. FPI is the number of non-mated probe searches that return a candidate at rank one with a score greater than a threshold, t. The false positives are not normalized by the proportion of non-mated searches. Otherwise, algorithms could configure their face detector to have many more false detections, thus lowering their FPIR. FNIR is the proportion of mated searches that do not return the mated gallery template at or above the same threshold t.

[7]The UAV captured 73 high resolution still images of participants as they proceeded through the Marketplace vignette.

	TDR (%)		**HR Only TDR (%)**	
	FDPI 10^{-1}	FDPI 10^{-2}	FDPI 10^{-1}	FDPI 10^{-2}
GOTS-1	2.4	0.9	9.7	1.8
MTCNN	11.5	2.7	27.6	3.1

Table 4. True detect rates (TDR) at operating points of 10^{-1} and 10^{-2} false detects per image (FDPI) for the benchmark algorithms. "HR Only" refers to the performance of the algorithm on only the high-resolution video from the Speco 04P30X cameras, examples of which are illustrated in Fig. 2(b)

4. Results

Baseline results for face detection in video, surveillance-to-single, and the surveillance-to-booking protocols are illustrated in this section. Due to space limitations, we do not provide benchmark results for the remaining 1:N protocols or provide IET performance. They will be provided with the database release.

4.1. Face Detection

Two baseline algorithms were evaluated using the IJB-S face detection protocol. First, a government-off-the-shelf (GOTS) algorithm was tested. This GOTS algorithm was designed specifically to detect faces in unconstrained imagery and is shown to be the top performing face detector in a recent face detection benchmark [5]. Secondly, we report performance from a TensorFlow implementation of a multi-task cascaded convolutional neural network (MTCNN) [26]. Results are presented in Table 4 for processing all videos and for a subset containing videos from only the higher resolution Speco cameras. Both detectors fail to achieve high detect rates. Performance on the subset with higher resolution video performs better as expected but still fails to achieve high detection rates indicating the difficult nature of this data. In either case, this may in part be explained by the fact that both algorithms were trained on image data from a different domain than surveillance video. Nevertheless, the MTCNN algorithm provides better detection performance in comparison to the GOTS detection al-

gorithm.

4.2. 1:N Identification

The Surveillance-to-Single and Surveillance-to-Booking identification protocols are end-to-end protocols designed to test joint detection and recognition performance. The primary difference between both protocols is the gallery. In Surveillance-to-Single, the gallery used only contains a single frontal still image. In contrast, Surveillance-to-Booking, contains frontal, $\pm 45^\circ$, $\pm 90^\circ$ yaw, and images collected with $\pm 20^\circ$ pitch angle.

We combine the MTCNN face detector with an implementation of Google's FaceNet[8], which was shown to achieve 98.7% accuracy on LFW [10] (labeled as FaceNet). Detected faces are encoded through FaceNet and the output is clustered using DBSCAN [6]. A single template is created from each identity cluster and then searched against the galleries. Additionally, we report performance for the following configurations of the FaceNet algorithm: (i) The MTCNN, DBSCAN, and FaceNet combination when the data is partitioned to include only videos from the higher resolution Speco cameras (labeled as FaceNet-Speco in subsequent plots), (ii) ground truth bounding boxes and FaceNet combination (FaceNet-GT), and (iii) ground truth bounding boxes with FaceNet on the partition of Speco only videos (FaceNet-GT-Speco).

Fig. 7 illustrates the EERR for the surveillance-to-single protocol. Clearly, the MTCCN and FaceNet algorithm combination fails to provide high retrieval rates through the first 50 ranks. This is not surprising given the poor detection performance of the MTCNN algorithm listed in Table 4. Performance on the subset of Speco videos is marginally higher, 7% (compared to 4%) at rank 5. A bigger increase is observed utilizing the ground truth detections. Specifically, rank 5 performance increases to 20%. With the combination of partitioning and ground truth bounding boxes, rank 5 performance increases up to 32%. Fig. 8 is a plot of the EERR for the surveillance-to-booking protocol. While EERR performance did increase slightly, it is very similar to the surveillance-to-single protocol suggesting that the additional subject media available in the gallery across different poses did not have a strong impact on performance.

5. Summary and Conclusion

We have introduced a new publicly available face dataset, the IARPA Janus Surveillance (IJB-S) Video Benchmark. Unlike media "in-the-Wild" datasets such as LFW [10], IJB-(A,B,C) [13, 24, 15], MS-Celeb-1M [9], and MegaFace [16], IJB-S focuses on unconstrained surveillance video, and includes over 350 surveillance videos spanning 30

[8]The implementation of the face recognizer can be found at https://github.com/davidsandberg/facenet

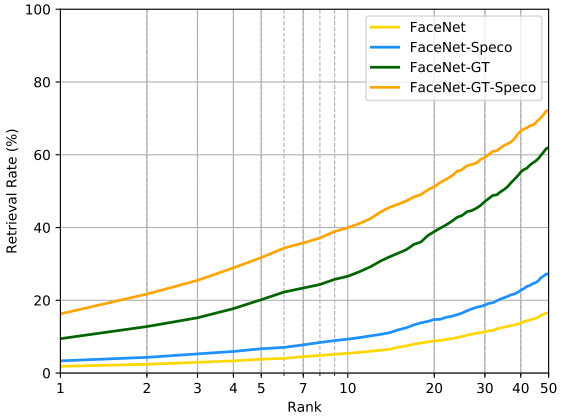


Figure 7. Average EERR performance across gallery sets G1 and G2 for the 1:N Surveillance-to-Single Identification protocol. The end-to-end (E2E) retrieval rate on the y-axis indicates the proportion of mated searches returned at or above a rank, incorporating misses from failed bounding box associations.

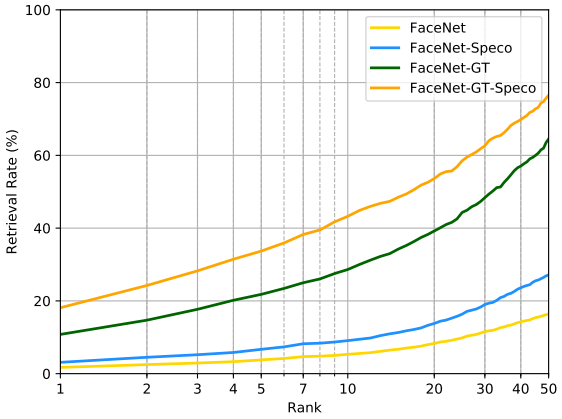


Figure 8. Average EERR performance across gallery sets G1 and G2 for the 1:N Surveillance-to-Booking Identification protocol.

hours, with over 10 million annotations, corresponding protocols, and baseline performance for face detection and open set identification experiments.

In this data set, each subject is present in at least 12 surveillance videos. In total, there are 5,656 enrollment images, and 404 enrollment videos. All still image and video media has manually annotated facial bounding boxes. The media within the dataset can be publicly redistributed through approved IRB collection plan. Along with the dataset, benchmark results from GOTS and an academic implementation of Google's FaceNet algorithm are released to be used for comparative research. Our benchmark results characterize the difficulty of the dataset, specifically, they highlight the importance of a robust face detector for generating subject tracks in surveillance video. The IJB-S dataset will be available through the NIST Face Projects website upon publication.

References

[1] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas. Face recognition from video: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(05):1266002, 2012.

[2] L. Best-Rowden, S. Bisht, J. C. Klontz, and A. K. Jain. Unconstrained face recognition: Establishing baseline human performance via crowdsourcing. In *IEEE IJCB*, pages 1–8, 2014.

[3] J. 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, et al. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE BTAS*, pages 1–8, 2013.

[4] A. Blanton, K. C. Allen, T. Miller, N. D. Kalka, and A. K. Jain. A comparison of human and automated face verification accuracy on unconstrained image sets. In *IEEE CVPR Workshop on Biometrics*, 2016.

[5] J. Cheney, B. Klein, A. K. Jain, and B. F. Klare. Unconstrained face detection: State of the art baseline and challenges. In *IEEE ICB*, pages 229–236, 2015.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

[7] M. Grgic, K. Delac, and S. Grgic. Scface — surveillance cameras face database. *Multimedia Tools Appl.*, 51(3):863–879, Feb. 2011.

[8] M. Günther, P. Hu, C. Herrmann, C. H. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kittler, et al. Unconstrained face detection and open-set face recognition challenge. *arXiv preprint arXiv:1708.02337*, 2017.

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. arxiv.org, August 2016.

[10] 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.

[11] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Transactions on Image Processing*, 24(12):5967–5981, 2015.

[12] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, 2008.

[13] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *IEEE CVPR*, pages 1931–1939, 2015.

[14] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, 99(3):303–331, 2005.

[15] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Cheney, and P. Grother. IARPA Janus benchmark – C: Face dataset and protocol. In *ICB*, 2018.

[16] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *IEEE CVPR*, 2017.

[17] J. C. Neves, G. Santos, S. Filipe, E. Grancho, S. Barra, F. Narducci, and H. Proença. Quis-campi: Extending in the wild biometric recognition to surveillance environments. In V. Murino, E. Puppo, D. Sona, M. Cristani, and C. Sansone, editors, *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, pages 59–68, Cham, 2015. Springer International Publishing.

[18] NIST. Special publication 500-271. Technical report, National Institute of Standards and Technology, 2007.

[19] P. J. Phillips et al. Overview of the multiple biometrics grand challenge. In *IEEE ICB*, pages 705–714, 2009.

[20] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[21] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin. Subsense: A universal change detection method with local adaptive sensitivity. *IEEE Transactions on Image Processing*, 24(1):359–373, Jan 2015.

[22] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[23] 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.

[24] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother. IARPA Janus Benchmark-B face dataset. In *IEEE CVPR Workshop on Biometrics*, July 2017.

[25] 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 *CVPR Workshop on Biometrics*, pages 81–88, June 2011.

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.