IARPA Janus Benchmark-B Face Dataset *

Cameron Whitelam † Emma Taborsky ‡ Austin Blanton Brianna Maze ‡
Jocelyn Adams ‡ Tim Miller ‡ Nathan Kalka ‡ Anil K. Jain § James A. Duncan ‡
Kristen Allen ¶ Jordan Cheney ‡ Patrick Grother ‖

Abstract

Despite the importance of rigorous testing data for evaluating face recognition algorithms, all major publicly available faces-in-the-wild datasets are constrained by the use of a commodity face detector, which limits, among other conditions, pose, occlusion, expression, and illumination variations. In 2015, the NIST IJB-A dataset, which consists of 500 subjects, was released to mitigate these constraints. However, the relatively low number of impostor and genuine matches per split in the IJB-A protocol limits the evaluation of an algorithm at operationally relevant assessment points. This paper builds upon IJB-A and introduces the IARPA Janus Benchmark-B (NIST IJB-B) dataset, a superset of IJB-A. IJB-B consists of 1,845 subjects with human-labeled ground truth face bounding boxes, eye/nose locations, and covariate metadata such as occlusion, facial hair, and skintone for 21,798 still images and 55,026 frames from 7,011 videos. IJB-B was also designed to have a more uniform geographic distribution of subjects across the globe than that of IJB-A. Test protocols for IJB-B represent operational use cases including access point identification, forensic quality media searches, surveillance video searches, and clustering. Finally, all images and videos in IJB-B are published under a Creative Commons distribution license and, therefore, can be freely distributed among the research community.

*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). 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.

†C. Whitelam is with Scientific Research Corporation, Atlanta, GA, U.S.A.

‡E. Taborsky, B. Maze, J. Adams, T. Miller, N. Kalka, J. Duncan, and J. Cheney are with Noblis, Reston, VA, U.S.A.

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

¶K. Allen is with Carnegie Mellon University, Pittsburgh, PA, U.S.A.

‖P. Grother is with the National Institute of Standards and Technology (NIST), Gaithersburg, MD, U.S.A.

1. Introduction

The development of scalable algorithms that can quickly and accurately recognize faces in unconstrained scenarios has become an overarching goal in computer vision. An "unconstrained" face recognition system should have the ability to perform successful face detection, verification and identification regardless of subject conditions (pose, expression, occlusion) or acquisition conditions (illumination, standoff, etc). Although automated systems are approaching human levels of face recognition on constrained imagery [16], there remains a large gap in unconstrained scenarios [4][3]. In order to close this gap, researchers must have access to large amounts of relevant training and testing data to design and evaluate their algorithms at operationally relevant assessment points (e.g., FAR of 0.1%). While several large training datasets have become available in the public domain [15], few unconstrained datasets have been released with operationally relevant protocols. The IARPA Janus Benchmark-B dataset, described here, contains a large annotated corpus of unconstrained face imagery that advances the state-of-the-art in unconstrained face recognition.

1.1. Unconstrained Face Imagery

The release of the "Labeled Faces in the Wild" (LFW) dataset in 2007 [7] spurred significant research activity in unconstrained face recognition. It was also influential in the release of subsequent unconstrained datasets such as the PubFig [10], YouTube Faces [18], MegaFace [13] and WIDER FACE [19] datasets. Extensive research in face recognition have now resulted in algorithms whose performance on the LFW benchmark has become saturated. Indeed, the best performance under the LFW protocol [15] exceeds 99% true accept rate at a 1.0% false accept rate.

One of the key limitations of the LFW, YTF, and PubFig datasets is that all faces contained in them can be detected by a commodity face detector, e.g. the Viola-Jones (V-J) face detector [17]. The V-J face detector was not designed to detect faces with a significant degree of yaw, pitch, or roll. As a result, available faces-in-the-wild datasets lack full pose variation. Another issue with the LFW dataset in

particular is the overlap of subjects in training and testing splits, which can result in training bias. To remedy this, the Benchmark of Large Scale Unconstrained Face Recognition (BLUFR) protocol was developed [11] as an alternative to the LFW protocol.

The MegaFace dataset was constructed by using the HeadHunter [12] algorithm for face detection. It includes one million images of 690K unique identities and is intended for use as a distractor set. WIDER FACE is a face detection benchmark dataset with 32,203 images and 393,703 annotated faces. Both of these datasets only have protocols designed for face detection, and thus cannot be used to evaluate face verification or identification directly.

The release of the NIST Face Challenge [6] and the IARPA Janus Benchmark A (IJB-A) dataset [9] in 2015 marked a milestone in unconstrained face recognition. Results from multiple submissions to the challenge show significantly inferior recognition performance compared to previous datasets. The best true accept rate at a 1.0% false accept rate is only 82.2% [6]). Although facial variations in IJB-A were not constrained by a commodity face detector, the dataset was hampered by a relatively small number of subjects (167 unique subjects) in each of the ten splits of the IJB-A protocol. This limited the ability to evaluate algorithms at lower ends of the ROC curve (e.g. FAR at 0.01% and 0.001%) due to the low number of possible impostor matches. Thus, in order to assess the accuracy and robustness of unconstrained face detection and recognition algorithms at more operationally relevant assessment points, a larger and more challenging unconstrained dataset needs to be available along with relevant evaluation protocols.

2. IJB-B Dataset

The IARPA Janus Benchmark-B (IJB-B) dataset contains Creative Commons licensed face imagery (i.e., open for re-distribution, provided proper attribution is made to the data creator) and contains both video and images. As an extension to the IJB-A dataset, face images and videos of 1,500 additional subjects were collected. After de-duplication with the publicly available VGG dataset [15] and the CASIA Webface dataset [20], 106 overlapping subjects were removed to keep the subjects in external training sets and IJB-B disjoint. An additional 49 subjects were removed due to erroneously or ambiguously annotated media after manual review. This resulted in a total of 1,345 subjects that were added to the existing 500 subjects in IJB-A, resulting in the IJB-B dataset consisting of 1,845 total subjects. All subjects are ensured to have at least two still images and one video in which their faces appear. All bounding boxes, eyes, and nose locations were manually annotated through Amazon Mechanical Turk, which is discussed in more detail below. Within IJB-B, there are a total of 21,798 (11,754 face and 10,044 non-face[1]) still images, with an average of ∼6 face images/subject, and 55,026 video frames pulled from 7,011 full motion videos, with an average of ∼30 frames/subject and ∼4 videos/subject. While the dataset is still biased towards subjects from North America and India, IJB-B contains a more uniform geographic distribution of subjects than that of IJB-A. Sample imagery from the IJB-B dataset can be seen in Fig. 1 while demographic and collection distributions can be seen in Figure 2. Table 1 describes key statistics of IJB-B and other unconstrained benchmark face datasets.

2.1. Collection Methodology

The collection and annotation methodology for all subjects in IJB-B is detailed below.

To determine viable and novel candidate subjects, 10,000 names were scraped from Freebase[2], a collaborative online database of real people. To ensure geographic and gender distribution, information for 500 women and 500 men

[1]Non-face imagery in IJB-B is used to evaluate face detection algorithms

[2]Since the time of collection and the submission of this paper, Freebase has transitioned to Wikimedia Commons

Dataset	# subjects	# images	#img/subj	# videos	# vid/subj	disjoint train/test set	pose variation
IJB-B	1,845	21,798	6.37	7,011	3.8	–	full
IJB-A [9]	500	5,712	11.4	2,085	4.2	–	full
LFW [7]	5,749	13,233	2.3	0	0	false	limited
YTF [18]	1,595	0	0	3,425	2.1	true	limited
PubFig [10]	200	58,797	294.0	0	0	true	limited
VGG [15]	2,622	982,803	375	0	0	–	limited
MegaFace [13]	–	1M	–	0	–	–	full
WIDER FACE [19]	–	32,203	–	0	–	true	full
CASIA Webface [20]	10,575	494,414	46.75	0	0	–	limited

Table 1: A comparison of IJB-B to other unconstrained face benchmark datasets. Full pose variation is defined as -90 to +90 degrees of yaw; anything less is regarded as limited pose variation. MegaFace and WIDER FACE are distractor and face detection sets, respectively, and as such do not contain subject labels.

Figure 1: Sample imagery from (a) IJB-B; (b) IJB-A; and (c) LFW [7] datasets. Note the challenging imagery that IJB-B contains, includes large facial occlusion, poor image quality, and greater illumination variation.

from 10 different geographic regions (10,000 subjects total) was collected. Once viable candidate names were determined, the number of available Creative Commons (CC)-licensed videos was determined for each candidate using the YouTube search API. Candidates were then sorted in descending order from the highest number of CC videos available to least.

Based on this list of potential subjects with CC videos, each candidate's name was automatically searched via Google and Yahoo images. For each subject, up to 20 Creative Commons images were downloaded for annotation via Amazon Mechanical Turk (AMT). If the presence of the subject was verified in at least 2 images, all CC YouTube videos associated with that subject were downloaded and released to AMT to confirm the subject's presence. If a subject was found to be in the video, AMT workers were then tasked to annotate start/stop times for subject's appearance in the video. These sections of the video were finally parsed into smaller clips and key frames were extracted using FFmpeg[3].

Each image and video key frame was first annotated with the bounding box locations of all faces. In turn, the bounding box for the person of interest was identified, and three fiducial landmarks (both eyes and nose base) were annotated. Metadata attributes such as occlusion, facial hair,

[3] www.ffmpeg.org

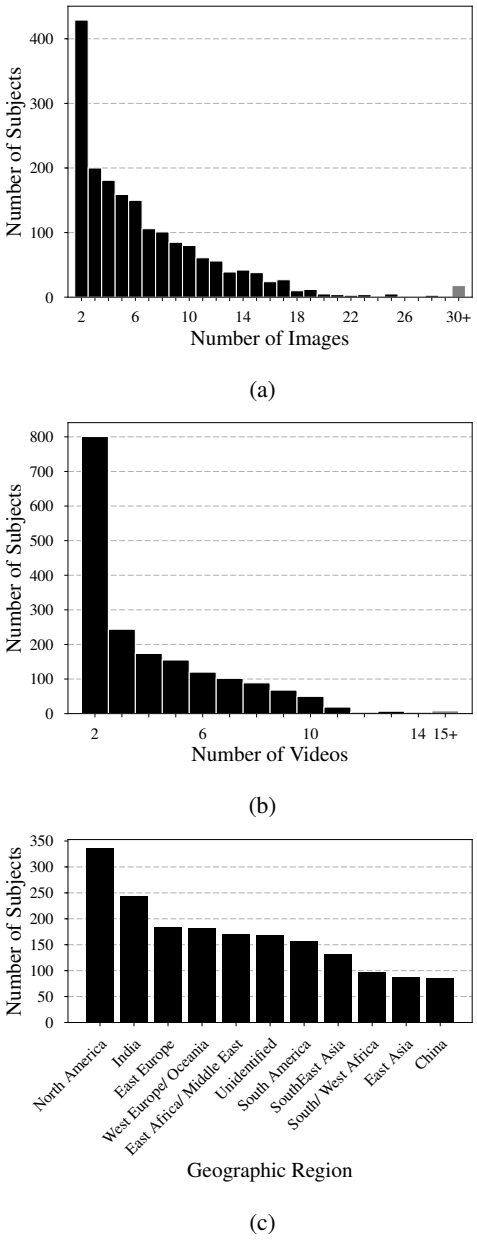

Figure 2: Geographic and media distribution in IJB-B. (a) number of images/subject; (b) number of videos/subject; (c) number of subjects/geographic location.

gender, etc. were also annotated. Finally, internal analysts inspected the annotated data to ensure correctness. An annotation example can be seen in Figure 3. For further details, please refer to [9].

In total, over 3.8 million manual annotations were performed by AMT workers. This annotation process resulted in (i) an accurate ground truth corpus of imagery containing bounding boxes facilitating face detection evaluations, (ii) subject labels for face recognition and clustering eval-

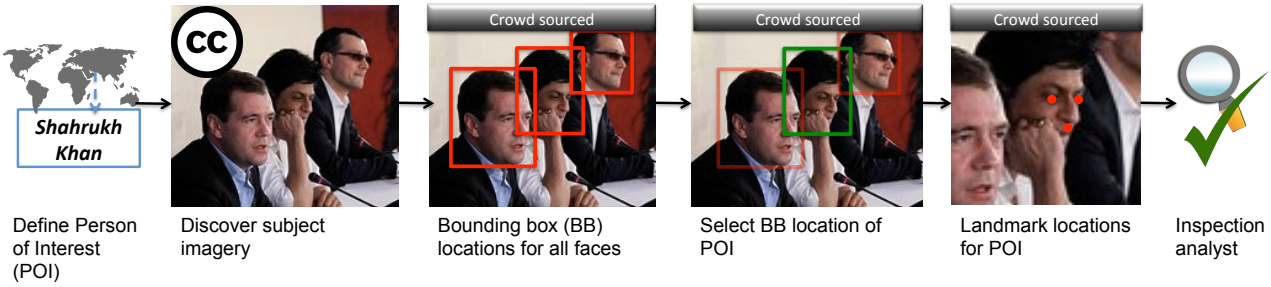

Figure 3: Overview of the data annotation process. Details can be found in [9].

uations, (iii) facial landmarks, allowing researchers to explore representation and learning schemes to handle full pose variations, and (iv) attribute metadata for understanding the effects different covariates (occlusion, facial hair, gender, capture environment, skintone, age, and face yaw) have on face recognition algorithm performance.

2.2. Dataset Contributions

The contributions of IJB-B are as follows:

- It is the largest annotated unconstrained joint detection, recognition, and clustering benchmark dataset released to date.
- The subject pool is the most geographically diverse of any public face dataset (including IJB-A) containing informative metadata annotations. Researchers can use these covariates to analyze the relative strengths and weaknesses of their algorithms.
- All media included in the dataset is available as a stable download that may be redistributed according to attribution and licensing information provided.
- The larger subject pool and the combination of video and imagery allows for more challenging and operationally relevant protocols.
- The IJB-B protocols will be released along with useful benchmark accuracy measures from a Government-Off-The-Shelf (GOTS) algorithm.

3. Protocol Description

The IJB-B dataset consists of 10 different protocols that test face detection, identification, verification, and clustering specified in the IARPA Janus program [1]. Due to space constraints, only 7 protocols are described below. A majority of the template designs created for IJB-B are based on the subject specific modeling described in [9]. In short, subject specific modeling refers to a single template being generated using some or all pieces of media associated with a subject instead of the traditional approach of creating multiple templates per subject, one per piece of media. This type of modeling produces a fixed length feature representation from any number of media pieces. This encourages, for example, temporal integration of face features from video clips. More traditional methods, where a feature vector is produced from each piece of media or from the single highest quality piece of media, is allowed as well. In all cases a single template is produced and used in recognition comparisons. Template size is one of the evaluation criteria and should be reported. Table 2 outlines some key statistics for all protocols developed thus far.

		Num. of Subjects	Num. of Templates	Pieces of Media
1:N Identification	**Gallery-S1**	931	931	3,081
	Gallery-S2	914	914	3,376
	Image	1,845	8,104	5,732
	Mixed	1,845	10,270	60,758
	Video	1,845	7,110	7,011
1:1 Verification	**Verify**	1,845	12,115	66,780
	Verify-Cov	1,845	68,195	66,780
Face Detection	**Face Partition**	$> 1,845$	N/A	66,780
	Non-face Partition	0	N/A	10,044
Clustering	**Clustering-32**	32	1,026	1,026
	Clustering-64	64	2,080	2,080
	Clustering-128	128	5,224	5,219
	Clustering-256	256	9,867	9,860
	Clustering-512	512	18,251	18,173
	Clustering-1024	1,024	36,575	36,092
	Clustering-1845	1,845	68,195	66,780

Table 2: Overview of the different protocols developed for IJB-B. Verify-Cov refers to Covariate Verification; Clustering-X denotes clustering with "X" number of subjects.

3.1. Face Detection

From the 76,824 (66,780 with faces and 10,044 without faces[4]) images/frames within IJB-B, AMT workers annotated a total of 125,474 faces or an average of $\sim$2 faces per piece of media. Typically, face detection evaluations assume that all images in the testing set contain a face. However, in operationally relevant use cases such as surveil-

[4]All non-face imagery used in the face detection protocol is public-domain, collected from Wikimedia Commons.

lance, the majority of media processed do not contain faces. Therefore, it is prudent to evaluate face detectors on non-face imagery that mirrors operational use cases. The IJB-B face detection protocol is augmented with 10,044 still images that do not contain any faces. Each non-face image was manually inspected to ensure that no faces were present.

3.1.1 Performance Metrics for Face Detection

Typically, face detection algorithms output confidence scores for every detected face. At a given threshold of False Detect Rate (FDR), the true detect rate (TDR) is measured as the fraction of ground truth faces correctly detected divided by the total number of faces in the ground truth. A correct detection is defined by a predicted bounding box which overlaps with at least 50% of the ground truth bounding box. The number of false accepts is defined as the number of reported detections that do not have an overlap of over 50% to any entry in the ground truth. If two or more boxes overlap by more than 50% with the ground truth bounding box, the box with the most overlap is considered the true detect while the other box or boxes are considered a false detect. Previous evaluations have reported this measure as total false accepts across a set of images [8], which limits cross dataset comparisons of face detection accuracy. Therefore, the False Detect Rate (FDR) should be measured as the number of false accepts divided by the number of images/frames tested [5]. Together, the TDR and FDR per image can be plotted in the form of a ROC, or listed at specific operating points. For the IJB-B face detection protocol, the evaluation reports TDR at a FDR of 0.1 and 0.01 (false detections of 1 in 10 images and 1 in 100 images). The average time to detect all faces in an image should also be reported as an extension metric.

3.2. Verification

The 1:1 verification protocol tests an algorithm's ability to match templates in an access control type scenario. Two separate verification protocols have been developed for IJB-B and are listed in more detail below.

3.2.1 1:1 Baseline Verification

A list of 8,010,270 comparisons is provided between S1 and S2 gallery templates and the 1:N Mixed Media probe templates (Section 3.2.2). In total, there are 10,270 genuine comparisons and 8,000,000 impostor comparisons. With the large numbers of genuine and impostor comparison scores, the lower bounds of the ROC at very low FAR values can be evaluated (0.01% or 0.001%). These values represent more operationally relevant operating points. This was not possible in IJB-A due to the relatively small number of genuine and impostor comparisons.

3.2.2 1:1 Covariate Verification

A second protocol was designed to test a face recognition algorithm's robustness on different covariates. In order to test for a specific covariate's influence, multiple gallery and probe images cannot be combined into a single template if they have differences within that covariate (e.g. one photo with a mustache and the other photo of the same subject with a clean shaven face). Therefore, only single image templates are used in this protocol. Each face image in IJB-B has been automatically labeled with estimated yaw using a GOTS pose-estimation algorithm and manually labeled with the following attributes: forehead occlusion (yes/no), eye occlusion (yes/no), nose/mouth occlusion (yes/no), gender (male/female), capture environment (indoor/outdoor), facial hair (none/beard/mustache/goatee), age group (0-19; 20-34; 35-49; 50-64; 65+; unknown), and skintone (1-6 increasing in darkness). In total, there are 20,270,277 comparisons (3,867,417 genuine and 16,402,860 impostor) between single image gallery and probe templates.

3.2.3 Performance Metrics for 1:1 Verification

For the 1:1 Verification protocols, the standard practice is to report the Receiver Operating Characteristic (ROC). Benchmark evaluations below report ROC metrics of TAR at a FAR of 1% and .01%. Also, the mean duration of template generation and comparisons should be reported as an extension metric.

3.3. Face Identification

The identification (1:N) protocol tests the accuracy of both open- and closed-set identifications using probe templates from all subjects on two disjoint galleries. Each gallery set, Set 1 (S1) and Set 2 (S2), contains a single template per subject that is created by using half of the still imagery media at random for that subject. For example, if a subject is associated with 12 still images and 32 video frames, 6 still images are chosen at random to create the gallery template. The remaining media instances are reserved for the probe set. Gallery sets S1 and S2 each initially contained 1,000 subjects. However, after de-duplication with VGG [15] and subject removal due to erroneous annotations, gallery sets S1 and S2 are composed of 931 and 914 subjects/templates using 3,081 and 3,376 still images, respectively. Both gallery sets are disjoint from each other so that open-set identification scenarios (i.e. probe templates that do not have a mate in the gallery) can be tested. In order to have a thorough understanding of an algorithm's strengths and weaknesses, multiple identification probe sets were designed to be tested with the S1 and S2 gallery sets. These probe set protocols are described in more detail below.

3.3.1 1:N Still Image

The 1:N Image Identification protocol was designed to test algorithms on their ability to recognize unconstrained faces in high quality, portrait style imagery. This protocol specifies multiple probe templates for each subject and uses only still image media. In total, there are 8,104 templates extracted from 5,732 still images. Probe templates are created using still image media that is disjoint from the subject's gallery template (in either S1 of S2), according to the following method:

- Each non-gallery image is used as a single image template.
- All non-gallery images are randomly divided in half, with each half being a multi-image template.
- All non-gallery images are used to construct a multi image template.

3.3.2 1:N Mixed Media

The 1:N Mixed Media Identification protocol was built to mimic forensic searches in which multiple types of media (including still images and video frames) are available to construct a single template. In total, there are 10,270 templates containing 60,758 still images and video frames. Probe templates are created using combinations of the subject's video frames and still images not present from the subject's gallery template. The probe template design methodology is as follows:

- For each video, all annotated frames in the video are concatenated to create a template.
- Half of the non-gallery still images are concatenated to create one template.
- The remaining half is concatenated with all video frame media for the subject to create another template.
- All non-gallery image media are combined to create a template.
- All frame media from all videos are combined to create a template.
- All non-gallery image media and all frame media are combined to create a template.

3.3.3 1:N Video

The 1:N Video Identification protocol was built to mimic video surveillance scenarios in which a subject of interest is to be tracked and recognized. This protocol specifies multiple probe templates for each subject and uses the first frame of the subject that is annotated as ground truth. In total, there are 7,110 probe templates for this protocol. Each subject is associated with at least one 1:N Video template.

3.3.4 Performance Metrics for 1:N Identification

The following two metrics are used: (i) Cumulative Match Characteristic (CMC) which measures the accuracy in a closed-set identification scenario. After returning a candidate list of subjects sorted by their matching score, the CMC is computed by calculating the percentage of probe matches that have a true-positive within the top k sorted returns (ranks) within the gallery; (ii) Identification Error Tradeoff (IET) characteristic expresses how the False Positive Identification Rate (FPIR) varies with respect to the False Negative Identification Rate (FNIR). FPIR is the proportion of probe searches that return one or more incorrect gallery candidates above a threshold t. FNIR is the proportion of probe searches that do not return the mated gallery template at or above the same threshold, t. While the CMC is used to measure closed set performance, the IET measures open-set 1:N performance, where probe searches are not guaranteed to have a mate in the gallery. Also, the mean compute time of template generation and probe searches should be reported as an extension metric.

3.4. Face Clustering by Identity

The clustering protocol was designed to test a face recognition algorithm's ability to identify multiple instances of the same subject from various pieces of media with no a priori subject labeling [14]. In total, there are seven sub-protocols that test an algorithm's ability to cluster at different scales. Details on the number of subjects and total pieces of media for each sub-protocol can be seen in Table 2. For each sub-protocol, all imagery for each selected subject in IJB-B is used (still images and video frames) and is a superset of the previous sub-protocol (e.g. Clustering-64 contains all media from Clustering-32 plus media from 32 unique new subjects). The input to the clustering protocol is an image and a bounding box that delineates the face of interest. This face is then treated as an item to be clustered. In addition, a hint is provided for each sub-protocol which serves as a rough order of magnitude on the number of subjects to be clustered and is calculated by $10^{\lceil \log_{10}(\#\text{ofsubjects}) \rceil}$. Benchmark evaluations in section 4.4 report the BCubed Precision and Recall [2] as well as F-measure for each clustering sub-protocol.

4. Baseline Experiments

Baseline accuracies for a majority of the protocols developed for IJB-B are listed below. These baselines are provided to allow researchers to evaluate whether their algorithms improve upon the government-off-the-shelf (GOTS) algorithm. Note that due to space, the 1:N Image Identification and 1:1 Covariate Verification results are not shown, but will be available from the authors.

4.1. Face Detection

Two baseline algorithms were used to test the IJB-B face detection protocol. First, a GOTS algorithm was used that was designed specifically to detect faces in unconstrained imagery. The GOTS detector was the top performing detector in a recent face detection benchmark [5], where it was shown to achieve results similar to top published performers on the FDDB dataset [8]. Secondly, the open sourced HeadHunter[5] algorithm was used which was also specifically designed for unconstrained imagery. The ROC for both baseline algorithms on the IJB-B face detection protocol can be seen in Figure 4. Notice that the GOTS and HeadHunter algorithms have similar performance, with the GOTS algorithm performing $\sim 10\%$ better than HeadHunter at a FDR of 0.01.

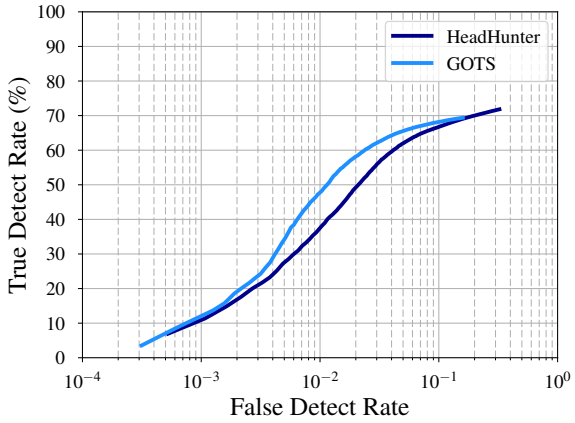

Figure 4: Face detection results on the IJB-B face detection protocol using the GOTS and HeadHunter algorithms. An overlap of 50% between the detected face and ground truth was used to determine True and False Detects.

4.2. Identification (1:N Mixed Media)

The 1:N Mixed Media Identification protocol was benchmarked on the GOTS algorithm discussed in Section 4.1 as well as one academic algorithm based on the Convolutional Neural Network described in [15]. The CNN benchmark used an open-source, state-of-the-art model trained on the VGG face dataset [15]. The averaging technique described in Algorithm 1 was applied to the CNN's output to process the multi-image IJB-B protocols. CMC and Identification Error Tradeoff characteristic results for these algorithms are shown in Figure 5 and 6, respectively.

[5] http://markusmathias.bitbucket.org/2014_eccv_face_detection/

Algorithm 1: Frame averaging

Data: Cropped faces from still images $\{I^S\}_{n=1}^N$ and from video frames $\{I^F\}_{k=1}^K$ belonging to the same subject $\mathcal{S}$

Result: Feature representation V

for $k \in \{1, \ldots, K\}$ **do**
  Propagate input $y_k^F \leftarrow \text{CNN}(I_k^F)$
end
for $n \in \{1, \ldots, N\}$ **do**
  Propagate input $y_n^S \leftarrow \text{CNN}(I_n^S)$
end
$V \leftarrow \frac{1}{N+1}\left[\mathbf{E}\left[y^F\right] + \sum_{n=1}^N y_n^S\right]$

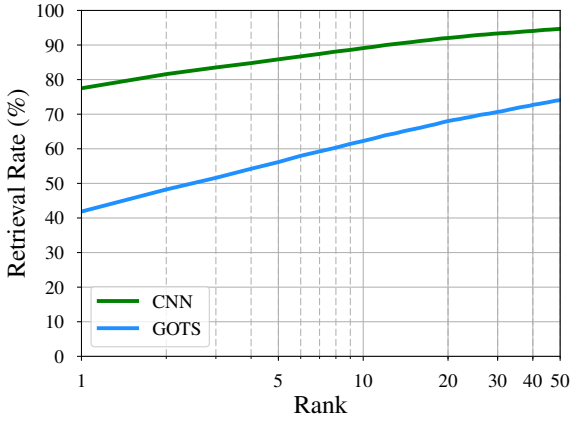

Figure 5: Average CMC performance across gallery sets S1 and S2 for the 1:N Mixed Media Identification protocol.

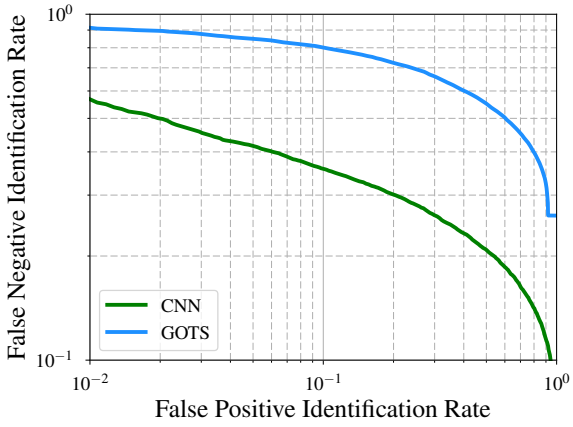

Figure 6: Average IET performance across gallery sets S1 and S2 for the 1:N Mixed Media Identification protocol.

	Hint	Precision	Recall	F-Measure	Run Time	Percent of FTEs
Clustering-32	100	0.589	0.298	0.395	0.32m	20.9
Clustering-64	100	0.578	0.302	0.396	1.15m	20.0
Clustering-128	1,000	0.605	0.352	0.445	7.23m	15.5
Clustering-256	1,000	0.581	0.362	0.446	25.12m	14.8
Clustering-512	1,000	0.516	0.328	0.401	76.92m	16.7
Clustering-1024	10,000	0.485	0.345	0.403	310.5m	15.5
Clustering-1845	10,000	NA				

Table 3: BCubed Precision, Recall, and F-measure values as well as the clustering hint (upper bound on no. of clusters), run time, and Failure To Enroll (FTE) rate for the GOTS algorithm on the IJB-B clustering sub-protocols. The number of images being clustered per sub-protocol can be found in Table 2. FTEs were ignored when calculating Precision, Recall and F-measure. Clustering-X indicates the number of ground truth clusters specified by X. Run times are reported using Intel® Core™ i7-6950X with 64GB of RAM.

4.3. Verification (1:1)

The algorithms discussed in Section 4.2 were used to create benchmark results for the 1:1 Verification protocol. These results can be see in Figure 7. Note that the CNN algorithm outperformed the GOTS algorithm at all ROC operating points.

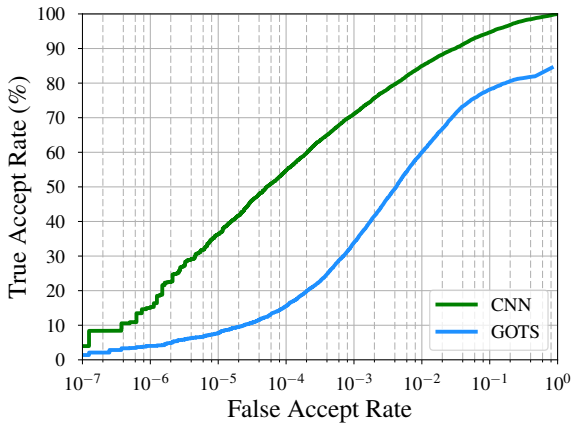

Figure 7: Average ROC performance across gallery sets S1 and S2 for the 1:1 Verification protocol.

4.4. Clustering

The GOTS algorithm described above was used to test all seven clustering sub-protocols. Table 3 shows the pairwise precision, recall, and F-measure as well as timing information and Failure to Enroll rates (FTEs). The number of images being clustered per sub-protocol can be seen in Table 2. Figure 8 shows sample clustering results. Note that all FTEs were ignored during evaluation. Note that the GOTS clustering algorithm was unable to execute the Clustering-1845 sub-protocol due to memory constraints.

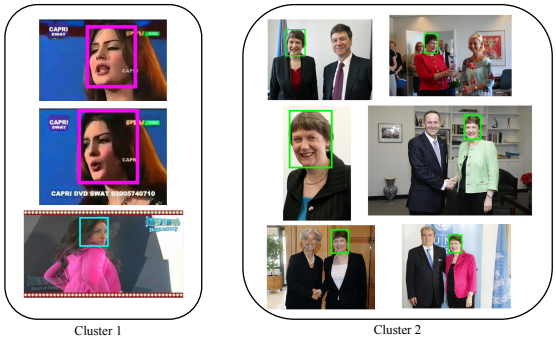

Figure 8: Sample clustering results using the GOTS algorithm on the Clustering-32 protocol. Unique subject identities are represented with different bounding box colors.

5. Summary

This paper has introduced the IARPA Janus Benchmark-B (IJB-B) dataset as an extension to the publicly available IJB-A dataset. IJB-B is an unconstrained face dataset that consists of still images, video frames, and full motion videos of 1,845 unique subjects. A total of 21,798 (11,754 with faces and 10,044 without faces) still images and 55,026 video frames pulled from 7,011 full motion videos have been manually annotated with ground truth face bounding boxes, landmarks and metadata (such as gender, facial hair, skintone, etc). IJB-B has been specifically designed to contain a larger number of diverse subjects, and more challenging operationally relevant protocols than that of IJB-A. All IJB-B media, including still images and videos, were collected under a Creative Commons distribution license and hence can be freely distributed among the research community. The protocols developed for IJB-B have been designed to test detection, identification, verification, and clustering of faces. IJB-B is the first unconstrained face dataset to provide specific clustering benchmark protocols. Benchmark results are presented from GOTS and an academic algorithm and can be used for comparative research. The IJB-B dataset will be available through the NIST Face Projects website[6].

[6] https://www.nist.gov/programs-projects/face-projects

References

[1] *IARPA Janus, IARPA-BAA-13-07*. 4

[2] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486, 2009. 6

[3] 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. 1

[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. 1

[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. 5, 7

[6] P. Grother and M. Ngan. The IJB-A face identification challenge performance report. Technical report, National Institute of Standards and Technology, 2016. 2

[7] 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. 1, 2, 3

[8] V. Jain and E. G. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, 2010. 5, 7

[9] 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. 2, 3, 4

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE ICCV*, Oct 2009. 1, 2

[11] S. Liao, Z. Lei, D. Yi, and S. Z. Li. A benchmark study of large-scale unconstrained face recognition. In *IEEE IJCB*, pages 1–8, 2014. 2

[12] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, pages 720–735. Springer, 2014. 2

[13] A. Nech and I. Kemelmacher-Shlizerman. Megaface 2: 672,057 identities for face recognition. 2016. 1, 2

[14] C. Otto, D. Wang, and A. K. Jain. Clustering millions of faces by identity. To appear in the *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2017. 6

[15] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 1, 2, 5, 7

[16] Y. Taigman, M. Yang, M. Ranzato, and L. . Deepface: Closing the gap to human-level performance in face verification. In *IEEE CVPR*, pages 1701–1708, 2014. 1

[17] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004. 1

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

[19] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[20] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. 2

IJB-B Image Attribution

Figure 1

- Jinmu Choi, "President Park Geun-hye, New Year's Day for Cultural Artists in 2014" is licensed under cc-by-3.0
- Magareto, "Tina Karol" is licensed under public domain
- Carlos t, "Mcvfelipecalderon" is licensed under public domain
- Carine06, "Flickr - Carine06 - Li Na" is licensed under cc-by-sa-2.0
- Lachezar, "Fidel Castro Ruz in Varna" is licensed under public domain
- lololol.net, "lololol - vol.1 / Jacques Ranciere is coming" is licensed under cc-by-3.0

Figure 7

- Pashto song Awo Ghazal, "GHAZALA JAVEED .pashto song" is licensed under cc-by-3.0
- Pashto song Awo Ghazal, "Avt Khyber 6 New Singer Ghazala Javeed ." is licensed under cc-by-3.0
- Khajanaa Boldndbeauty, "Rituparna Sengupta" is licensed under cc-by-3.0
- Maureen Lynch, "Helen Clark and Jeffrey Sachs" is licensed under cc-by-2.0
- United Nations Development Programme, "Helen Clark and Ulla Toernaes, UNDP" is licensed cc-by-2.0
- Maureen Lynch/UNDP, "Helen Clark by Maureen Lynch" is licensed under cc-by-2.0
- United Nations Development Programme, "John Key Helen Clark handshake" is licensed under cc-by-2.0
- United Nations Development Programme, "Helen Clark in France" is licensed under cc-by-2.0
- United Nations Development Programme, "Helen Clark and Sali Berisha, UNDP" is licensed under cc-by-2.0