To Appear: The Twenty-Eighth IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, U.S.A., June 8-10, 2015.

Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A *

Brendan F. Klare† Ben Klein† Emma Taborsky† Austin Blanton† Jordan Cheney†
Kristen Allen† Patrick Grother‡ Alan Mah§ Mark Burge¶ Anil K. Jain‖

Abstract

Rapid progress in unconstrained face recognition has resulted in a saturation in recognition accuracy for current benchmark datasets. While important for early progress, a chief limitation in most benchmark datasets is the use of a commodity face detector to select face imagery. The implication of this strategy is restricted variations in face pose and other confounding factors. This paper introduces the IARPA Janus Benchmark A (IJB-A), a publicly available media in the wild dataset containing 500 subjects with manually localized face images. Key features of the IJB-A dataset are: (i) full pose variation, (ii) joint use for face recognition and face detection benchmarking, (iii) a mix of images and videos, (iv) wider geographic variation of subjects, (v) protocols supporting both open-set identification (1:N search) and verification (1:1 comparison), (vi) an optional protocol that allows modeling of gallery subjects, and (vii) ground truth eye and nose locations. The dataset has been developed using 1,501,267 million crowd sourced annotations. Baseline accuracies for both face detection and face recognition from commercial and open source algorithms demonstrate the challenge offered by this new unconstrained benchmark.

*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.

†B. Klare, B. Klein, E. Taborsky, A. Blanton and J. Cheney are with Noblis, Falls Church, VA, U.S.A.

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

§A. Mah is a consultant, Washington, DC, U.S.A.

¶M. Burge is with the Intelligence Advanced Research Projects Activity (IARPA), McLean, VA, U.S.A.

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

1. Introduction

The development of accurate and scalable unconstrained face recognition algorithms is a long term goal of the biometrics and computer vision communities. The term "unconstrained" implies a system can perform successful identifications regardless of face image capture presentation (illumination, sensor, compression) or subject conditions (facial pose, expression, occlusion). While automatic, as well as human, face identification in certain scenarios may forever be elusive, such as when a face is heavily occluded [9] or captured at very low resolutions, there still remains a large gap between automated systems and human performance on familiar faces [15]. In order to close this gap, large annotated sets of imagery are needed that are representative of the end goals of unconstrained face recognition. This will help continue to push the frontiers of unconstrained face detection and recognition, which are the primary goals of the IARPA Janus program [1].

1.1. Unconstrained Imagery

A key step towards advancing unconstrained face recognition was the release of the "Labeled Faces in the Wild (LFW)" dataset in 2007 [6]. This dataset contains still images of subjects captured in non-controlled, or "wild", settings, downloaded from the web. Early recognition rates on this dataset were quite low. With an unprecedented lack of constraint on face image capture at that time, LFW was dubbed by many as the first unconstrained face recognition dataset. Since the inception of LFW, many similar datasets have been released, including but not limited to PubFig [11], and YouTube Faces (YTF) [19].

The release of LFW spurred significant activity by researchers in academia, major face recognition vendors, and software technology companies. As a result, performance has begun to saturate on LFW, YTW, and other unconstrained datasets [17]. At the same time, unconstrained face recognition is hardly considered a solved problem, as supported by poor performance of state of the art face recognition algorithms on surveillance videos. This is partially attributed to protocols that do not capture requirements of

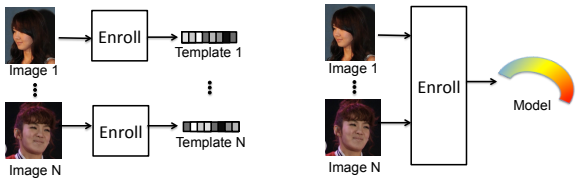
	(a)	(b)	(c)
Types of Images:	Frontal, Cooperative subject, Controlled environment	Near frontal, uncooperative, minimal environment variations (e.g., LFW)	Full variation in pose, illumination, environment
Automated detection ability:	Human performance	Near human performance	Cannot detect consistently
Automated recognition ability:	Human performance	Near human performance	Cannot recognize

Figure 1. (a) Face recognition accuracy on frontal images is considered to be a nearly solved problem. (b) Accuracy has greatly improved on face images captured in unconstrained settings that can be detected with a commodity face detector. (c) Example of an image in the IARPA Janus Benchmark A; this is believed to be the most unconstrained face dataset to date.

many operational unconstrained scenarios [12]. However, a more apt explanation for why unconstrained face recognition is still far from being a solved problem is that datasets such as LFW are not fully unconstrained. Instead, a key limitation exists with these datasets, namely: *a commodity face detector was used to detect all the faces included in the database*. This use of a commodity face detector constrains, among others, allowable pose variation, occlusions, and illuminations conditions.

The commodity face detector used to collect many of the aforementioned datasets is OpenCV's Viola Jones face detector [18]. Despite the many desirable properties of this detector (it is scalable, accurate on frontal faces, and available in the open source), it is not designed to detect faces with full pose variation. Instead, the detector is trained to detect near frontal faces. As such, a key limitation exists in the available public domain datasets: the faces lack full pose variation. While significant progress is also being made in unconstrained face detection, a gap still exists between algorithms that can detect fully unconstrained faces at low false positive rates, and the capability of face recognition algorithms [13, 4].

In summary, the current state of the art in unconstrained face recognition is high accuracy (roughly 99% true accept rate at a false accept rate of 1.0% [17]) on faces that can be detected with a commodity detectors, but unknown accuracy on other faces. Despite the fact that face detection and recognition research generally has advanced somewhat independently, the frontal face detector filtering approach used for key in the wild face recognition datasets means that progress in face recognition is currently hampered by progress in face detection. Hence, a major need exists for a face recognition dataset that captures as wide of a range of variations as possible to offer challenges to both face detection as well as face recognition.



(a) (b)

Figure 2. Commercial face recognition systems currently operate by generating a template for each input image, even when multiple images of a subject are available, as illustrated in (a). An alternate paradigm is to leverage multiple images of a subject at enrollment time to learn a subject specific model, as illustrated in (b). The IARPA Janus Baseline A recognition protocol makes a distinction between these two approaches, as they offer an important tradeoff in accuracy and efficiency.

1.2. Subject Specific Modelling

Commercial face recognition technology operates in a template-based manner: an input image results in a proprietary template (i.e., feature vector) representing the face which is stored internally in the database. If multiple images of a subject are available at enrollment time, then multiple templates will be stored. This paradigm has been important for achieving scalable systems, and the recognition accuracy when employing this technique on controlled capture imagery has been demonstrated to be quite high [5].

As we seek to close the gap in face recognition accuracy on unconstrained imagery, a shift in this paradigm must be considered. For example, humans are often noted for our remarkable ability to perform face recognition in challenging scenarios. However, there is a large difference between our performance on familiar and unfamiliar faces [14]. This divergence suggests some degree of subject specific representations in the human brain.

Many of the most high profile scenarios for deploying face recognition technology involves watch list scenarios for persons of interest. In many of these cases, we have several images of a person of interest. For example, the FBI's most wanted list[1] has several images for each person of interest. Similarly, video imagery from subjects often contains several different viewpoints.

In order to fully exhaust available approaches to unconstrained face recognition, subject-specific modeling must be explored (as illustrated in Figure 2). This implies that training could occur on an active gallery to learn the nuances of subjects that are labelled in a gallery. While this approach is believed to more closely align automated technology with human cognition, it comes with potentially severe implications for designing scalable systems. That is, if an

[1] www.fbi.gov/wanted

(a) (b) (c) (d) (e)

Figure 3. Examples of the faces in the IJB-A dataset. These images and video frames highlight many of the key characteristics of this publicly available dataset, including full pose variation, a mixture of images and videos, and a wide variation in imaging conditions and geographic origin.

algorithm is developed to perform subject specific modeling (e.g., training per-subject SVM's), then image and video enrollment speeds are no longer near-constant as they will also involve training. Further, multiple labelled images of a subject are not always available for enrollment, so such techniques may not always be relevant. Regardless, in order to properly explore all approaches to unconstrained face recognition, protocols must both allow, and delineate, difference between template and model-based approaches.

Along similar lines is the distinction between still images and videos of a subject. The majority of unconstrained databases available operate on only still images (e.g., LFW, PubFig) or videos (e.g., YTF). However, in practice both may be available [2]. As template-based and model-based algorithms are explored, the benefit of having either still images or videos becomes important to study. For example, a given model-based solution may work particularly well on videos as it can explicitly leverage temporal information in the multiple frames available. At the same time, such an approach may struggle on still images as only few uncalibrated images may be available. Conversely, template-based methods may not perform well on video data as the abundance of samples cannot be explicitly leveraged.

For the remainder of the paper, in order to maintain consistency with existing nomenclature for face recognition, both subject-specific models and templates generated from single pieces of imagery with be referred to as "templates".

1.3. Paper Organization

The remainder of the paper is organized as follows. In Section 2 details of the proposed dataset are provided. Section 3 discusses face recognition using the proposed dataset, to include an overview of the protocols, operational analogies for these protocols, and baseline results. Section 4 discusses face detection within the proposed dataset, to include an overview of the face detection protocol, and baseline detection accuracies from multiple off the shelf detectors. This includes an overview of the provided protocols and baseline accuracies from leading commercial off the shelf face recognition algorithms.

2. Proposed Dataset

In this paper we introduce the IARPA Janus Benchmark A (IJB-A)[2], which is publicly available for download[3]. The IJB-A contains images and videos from 500 subjects captured from "in the wild" environment. All labelled subjects have been manually localized with bounding boxes for face detection, as well as fiducial landmarks for the center of the two eyes (if visible) and base of the nose. Manual bounding box annotations for all non-labelled subjects (i.e., other persons captured in the imagery) have been captured as well. All imagery is Creative Commons licensed, which is a license that allows open re-distribution provided proper attribution is made to the data creator. The subjects have been intentionally sampled to contain wider geographic distribution than previous datasets. Recognition and detection protocols are provided which are motivated by operational deployments of face recognition systems. An example of images and video from IJB-A can be found in Figure 3.

2.1. Collection Methodology

A significant amount of effort is required to collect, annotate and verify such a large corpus of imagery. The procedure for collection and annotation described in this section was motivated by the need for a repeatable and scalable workflow.

Each subject in the data corpus was manually specified (e.g., Japanese Prime Minister, Shinzo Abe); this specification procedure was performed such that geographic origin of subjects were generally well distributed across the globe. Once a subject was specified, images and videos of the subject were located by performing internet searches on Creative Commons licensed imagery. For each identified

[2]As the IARPA Janus program continues, additional datasets may be provided in the public domain.

[3]That IJB-A dataset, protocol files, and benchmark leader boards are available at: http://face.nist.gov

image or video clip, the subject name, url, and (if a video) start and stop time of the subject's appearance were stored in a spreadsheet. At the end of each day, automated scrapping software downloaded the subject's imagery and stored all relevant information in a relational database. For videos, a clipped version from the original codec was stored, as well as the extracted I-frames.

After curating imagery, the next step was annotation by utilizing the crowdsourcing service, Amazon Mechanical Turk (AMT), which consisted of three distinct tasks. The first annotation task was annotating a bounding box around all faces in an image or video frame. Specific visual and written guidance was given to annotators to place the bounding box around the boundary of the head. Each image was annotated by at least five AMT workers. In order to consolidate the multiple annotations into a single set of annotations, the following approach was applied.

Once all faces in an image/video frame were annotated with bounding boxes, the next step was to determine which bounding box corresponded to the person of interest. This task is needed because while it was known that the person of interest was in the image, given multiple persons in many of the downloaded images, we needed to confirm which belonged to the person of interest. To support this task, we collected one good quality reference image for each subject. The annotators were then shown the reference image alongside the target image and asked to select the bounding box that corresponded to the person of interest.

At this point the location of the person of interest was known. While recognition could be performed using the bounding box information alone, more specific face localization information was desired given the difficult nature of the imagery. As such, the location of the center of each eye (if visible) and the base of the nose were annotated for the person of interest in each image and video. Given the wide range of pose variation, many images and frames exist with only one eye visible.

Finally, manual inspection was performed on all subjects to verify the correctness of the consolidated annotations. In cases where annotations were erroneous, the information was manually rectified by a well informed (non-crowd sourced) analyst. More a more comprehensive overview of the annotation process, readers are referred to [16].

The result of this annotation process consists of: (i) an accurate ground truthed corpus of imagery containing bounding box locations for all faces, thus facilitating face detection evaluations, (ii) subject labels, thus facilitating face recognition, and (iii) facial landmarks, thus allowing researchers to explore representations and learning schemes while improved landmark detection algorithms are developed to handle full pose variations. In total, 1,501,267 manual annotations were performed to prepare the IJB-A dataset. Additional meta-data was collected specifying the subject's skin color and gender, and, for each image and video frame, the facial pose, occlusion (eyes, mouth/nose, forehead), and environment (indoor or outdoor) [8].

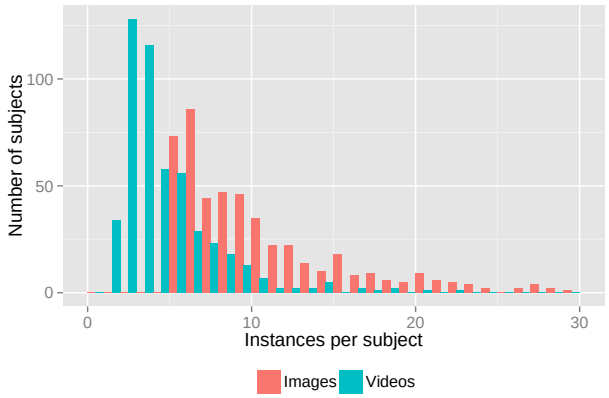


Figure 4. The IJB-A dataset contains a mix of images and videos for 500 labelled subjects. Shown are distributions of the number of images and videos per subject.

2.2. Perceived Contributions

The implication of the aforementioned data collection strategy for the IJB-A dataset has the following claimed contributions: (i) The most unconstrained database released to date; (ii) The first joint face detection and face recognition benchmark dataset collected in the wild; (iii) Meta-data providing subject gender and skin color, and occlusion (eyes, mouth/nose, and forehead), facial hear, and coarse pose information for each imagery instance; (iv) Widest geographic distribution of any public face dataset; (v) The first in the wild dataset to contain a mixture of images and videos; (vi) Clear authority for re-distribution; (vii) Protocols for identification (search) and verification (compare);

Table 1. A comparison of key statistics of the proposed IJB-A dataset and seminal unconstrained face recognition datasets.

Dataset	# subjects	# images	# img/subj	# videos	# vid/subj	sensor	environment	pose variation
IJB-A	500	5,712	11.4	2,085	4.2	varied	varied	full
LFW [6]	5,749	13,233	2.3	0	0	varied	varied	limited
PubFig [11]	200	58,797	294.0	0	0	varied	varied	limited
YTF [19]	1,595	0	0	3,425	2.1	varied	varied	limited
PaSC [3]	293	9,376	32	2,802	9.6	consistent	consistent	full

(viii) Baseline accuracies from off the shelf detectors and recognition algorithms; and (ix) Protocols for both template and model-based face recognition.

Every subject in the dataset contains at least five images and one video. IJB-A consists of a total of 5,712 images and 2,085 videos, with an average of 11.4 images and 4.2 videos per subject.

3. Face Recognition

3.1. Applications and Metrics

The IJB-A protocol focuses on two primary face recognition applications: the ability to search for a person's face in a set of images (*search*), and (ii) the ability to compare facial imagery from two persons, and verify whether or not they are the same person (*compare*). For both scenarios, the metrics set forth will measure both type I error (false positives) and type II error (false negatives).

Search Face searching is relevant to forensic identification, watch list identification, and de-duplication (e.g., mug shot repositories, ID card databases). The goal of face search is to determine whether a subject exists in a database of face images and/or videos (commonly referred to as the gallery). The search is conducted using a template generated from query images and/or videos (commonly referred to as the probe).

The definition of "accurate" facial search is application dependent. For user driven searches (e.g., forensic identification), accuracy is based on: (i) the number k of rank-ordered candidate retrievals results a user would examine for a typical query, and (ii) what percentage of probe searches return the probe's gallery mate within the top k rank-ordered results. The cumulative match characteristic (CMC) generally captures this scenario, as it reports the percentage of probes identified within a given rank (the independent variable).

For large scale, high throughput face search systems (e.g., de-duplication, watch list identification), not all searches can be manually examined. Instead, manual adjudication of search results will only occur if an application-dependent match score threshold is exceeded. In these cases, there is a trade off between false alarms, which expend system resources (manual adjudicators), and misses, which undermine the intent of the system (identifying whether a subject exists in a database). At a given threshold t, the false alarm rate (i.e., the false positive identification rate (FPIR), or the type I error rate) measures what fraction of comparisons between probe templates and non-mate gallery templates result in a match score exceeding t. This is tantamount to resources required to operate a system. The miss rate (i.e., the false negative identification rate (FNIR), or the type II error rate) measures what fraction of probe searches will fail to match a mated gallery template above a score of t. This is representative of how often the system fails to identify a subject. For a given matcher and database, there is only one degree of freedom between t (the independent variable), $FNIR$, and $FPIR$. These statistics are generally plotted in a decision error tradeoff (DET) characteristic. Mathematical expressions for CMC, FNIR, and FPIR can be found in [5].

Table 2. Geographic distribution of subjects contained in IJB-A.

Continent	# of subjects
Asia	89
Oceania	7
North America	135
South America	50

Continent	# of subjects
Europe	149
Middle East	29
Africa	41

Compare Facial comparison, or verification, is relevant to access control systems, re-identification, and application-independent evaluations of face recognition algorithms. Face recognition accuracy for verification is classically measured using the receiver operating characteristic (ROC). At a given threshold (the independent variable), ROC analysis measures the true accept rate (TAR), which is the fraction of genuine comparisons that correctly exceed the threshold, and the false accept rate (FAR), which is the fraction of impostor comparisons that incorrectly exceed the threshold.

3.2. Protocols

This section specifies the approach for generating the *search* and *compare* protocol files provided in the IJB-A database. An overview of these protocols are provided in Figure 6 Researchers are encouraged to process these protocols and plot corresponding accuracies using the provided evaluation harness[4].

Separate protocols are provided for search and compare. The following specifications pertain to both the search and compare protocols. There are ten random training and testing splits which occur at subject level, using all 500 IJB-A subjects. For each split, 333 subjects are randomly sampled and placed in the training split. These subjects are available for algorithms to build models and learn the variations in facial appearance that are representative of the Janus challenge. The remaining 167 subjects are placed in the testing split. Additional imagery may be used to train an algorithm under the strict condition that no such imagery contain the same subjects that are in the test split[5]. Bootstrap samples of training and testing splits are performed instead of cross validation to increase the number of testing subjects.

[4] http://libjanus.org

[5] There is a partial overlap of subjects in LFW and YTF, and IJB-A. The subjects are identified in the IJB-A dataset

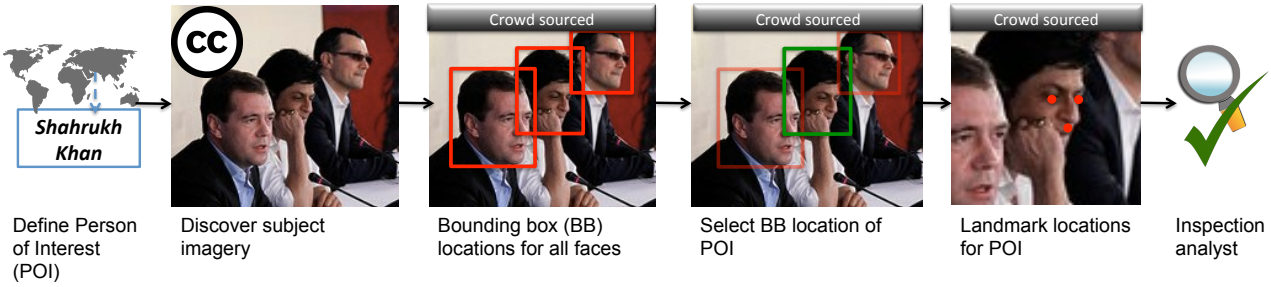


Figure 5. Overview of the data collection and annotation process. The first step involved selecting subjects, or "persons of interest", that, in aggregate, have wide geographic distribution. For each subject, Creative Commons (CC) licensed images and videos were discovered and ingested. Using crowd sourced labor, multiple annotations were performed for each image and video I-frame for the bounding box location 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. Finally, the analysts inspected the data to ensure correctness.

For each split, every testing subject has their imagery randomly sampled into either the probe set or the gallery set. The gallery set represents imagery contained in an operational database (e.g., mug shot repositories, databases for access control systems). The probe set represents imagery used to query a search system, or to compare against a specified ID in a verification system. Each test subject $i, 1 \leq i \leq 167$, has n_i total images and videos instances, where $n_i \geq 6$ because every subject has at least five images and one video. $n_i^g, 1 < n_i^g < n_i$, of these instances are randomly sampled into the gallery set. The remaining instances $n_i^p, n_i^p + n_i^g = n_i$, are allocated to the probe set. The protocol specifies multiple probe templates for each subject. Specifically, n_i^p probe templates exist, corresponding to each n_i^p of the images and videos for subject i. Another probe template exists, which contains all n_i^p images and videos. If $n_i^p > 2$, two final probe templates are specified, each built with n_i^{p1} and n_i^{p2} random pieces of imagery, which are randomly selected from the probe set such that: $n_i^{p1} + n_i^{p2} = n_i^p$, the probe set imagery used in each template are non-overlapping. Probe templates are specified in this manner to facilitate analysis of different scenarios for querying a face recognition system. In many cases, only a single image or video exists. However, scenarios such as forensic search often involve multiple probe samples, which this protocol represents with the multi-instance probe samples.

Search The search protocol measures the accuracy of open-set and closed-set search on the gallery templates using probe templates. To prevent an algorithm from leveraging apriori knowledge that every probe subject contains a mate in the gallery [5], 55 randomly selected subjects in each split have templates/imagery removed from the gallery set. Every probe template in a given split (regardless of whether or not the gallery contains the probe's mated templates) are to be searched against the set of gallery templates. For each search, at a minimum, the following information is recorded: the id's of the 30 closest matching gallery templates, and the corresponding match scores to these 30 templates. Using these results, the following accuracy metrics are to be reported: the rank-1 and rank-5 accuracy, and the miss rate corresponding to false alarm rates of 1 / 10 and 1/100. CMC and DET characteristics should also be plotted where space permits.

Compare The compare protocol measures the verification accuracy. In a similar manner as previous benchmarks [6, 19], the protocol specifies precisely which genuine and impostor comparisons should be performed for each split. For a given split, the number of genuine comparisons will be equal to the number of probe templates, as a single gallery template exists for each subject. The number of impostor comparisons is set to 10,000. The impostor comparisons are randomly sampled between probe templates and non-mated gallery templates under the following restriction: the two subjects represented in the gallery and probe templates have the same gender, and their skin color differs by no more than one of the six possible levels. This approach is meant to garner more challenging impostor comparisons. Each of the specified genuine and impostors comparisons are to be conducted, and the corresponding match score is to be recorded. Using these results, the following accuracy metrics are to be reported: the TAR at

Protocol	Applications	Accuracy Metrics
Compare	1:1 match; Access control; Re-identification	TAR @ FAR of 0.1, 0.01, and 0.001; ROC plot (TAR vs. FAR)
Search	De-duplication; Watch list; Forensic	FNIR @ FPIR of 0.1 and 0.01; Rank 1 and 5 accuracy; CMC plot; DET plot (FNIR vs. FPIR)

Figure 6. Overview of the IJB-A recognition protocols. The search applications are measured using open-set identification.

a FAR of 1 / 100 and 1/ 1000. The ROC should be plotted where space permits.

3.3. Reporting results

Statistics The mean and standard deviations for the following statistics must be reported for valid benchmark participation: (i) the accuracy metrics provided in the protocol (mean and s.d. measured across splits), (ii) enrollment duration, (iii) comparison duration, (iv) search duration, and (v) template size. All of these are critical factors in deployed systems. Reporting these metrics helps distinguish which applications are relevant to the corresponding algorithm.

Conditions The following conditions must be reported for valid benchmark participation: (i) whether or not the statistical learning was performed on the gallery, (ii) whether or not the manually provided landmarks were used to generate templates, and (iii) a statement that the protocol was precisely followed. An online leaderboard will distinguish between methods that implement gallery training, and those that do not.

3.4. Baseline

Baseline accuracies for the IJB-A were generated using a government-off-the-shelf (GOTS) algorithm, and the open source face recognition algorithm OpenBR [10]. The GOTS algorithm was specifically designed for unconstrained face recognition, and is competitive with top algorithms in the literature. This baseline allows researchers to understand whether or not their methods improve over operationally deployable technology. The OpenBR algorithm uses standard practices in face recognition algorithms, and is freely available in the open source[6].

Figure 7 contains CMC, DET and ROC plots using the previously described protocol. Table 3 contains the same results at specified operating points. The accuracies demonstrate the difficulty of the provided benchmark dataset. The accuracies in many cases are improved because the protocols contain both probe and gallery templates that consist of multiple images and/or videos, thus replicating many operational scenarios where such information is available. Other subjects only contain single instance, as is often the case as well. Many of the images failed to enroll for the algorithms, as only a single eye was visible: in such cases match scores are set to the minimum value. The accuracies provided are "without manual landmarks". Metrics for durations and template sizes will be on the IJB-A website.

[6] http://openbiometrics.org

4. Face Detection

4.1. Protocol

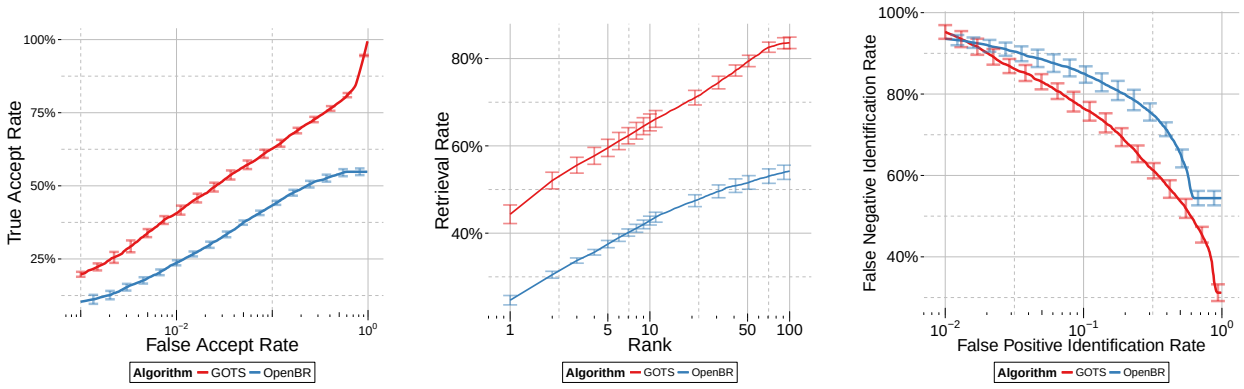
Face detection is a more straightforward application to experimentally evaluate than face recognition. The goal of face detection is to find all faces present in an image, and not falsely detect non-faces. Most face recognition algorithms output a confidence threshold for a given detection. When processing a set of images and/or videos using a face detector, results are collected in the form the location of reported detections, and the corresponding confidences. In turn, at a given threshold (the independent variable), the true detect rate (TDR) is measured as the fraction of ground truth faces correctly detected. Correct detection is defined by a predicted bounding box overlapping with at least 50% of the ground truth bounding box. The number of false accepts is the number of reported detections that do not correspond to any entry in the ground truth. While previous databases have reported this measure as a total false accepts across a set of images [7], this limits cross database comparisons of face detection accuracy. Instead, we measure false accepts as false detect rates (FDR) per image [4]. That is, the number of false accepts divided by the number of images/frames tested. Together, the TDR and FDR per image can be plotted in the form of a ROC, or listed at specific operating points, such as 1 false detection per 100 images (i.e, a FDR of 0.01).

The locations of all faces in all images and videos in IJB-A have been manually ground truthed by human annotators. In total, there are 67,183 faces of which 13,741 are from images and 53,442 are from videos. The imagery in the dataset has been partitioned into 10 randomly sampled training and testing sets, with two thirds of the imagery made available for training detector, and the remaining one third of the imagery used for testing. These are the same splits used in the recognition protocol. All face bounding boxes are 36 pixels or larger. While ground truth annotations were collected for smaller sized faces, they were removed because of a lack of identifiable information at such resolutions.

The accuracy metrics for the IJB-A face detection results are the TDR at a FDR per image rate of 1 / 10, and 1 /100.

Table 3. Specific recognition accuracies required to be reported on the IJB-A dataset. Results are from the baseline GOTS algorithm, and the open source algorithm OpenBR.

	TAR @ FAR's of:			CMC:		FNIR @ FPIR's of:	
Algorithm	0.1	0.01	0.001	Rank-1	Rank-5	0.1	0.01
GOTS	0.627 ± 0.012	0.406 ± 0.014	0.198 ± 0.008	0.443 ± 0.021	0.595 ± 0.02	0.765 ± 0.033	0.953 ± 0.0236
OpenBR	0.433 ± 0.006	0.236 ± 0.009	0.104 ± 0.014	0.246 ± 0.011	0.375 ± 0.008	0.851 ± 0.028	0.934 ± 0.017



(a) (b) (c)

Figure 7. Benchmark recognition accuracies on the proposed IJB-A dataset using a GOTS face recognition algorithm. (a) ROC plot for the compare protocol. (b) CMC plot for the search protocol. (c) DET plot for the search protocol.

When not limited by space, researchers should also plot the ROC and precision/recall curves. Algorithms must consider a minimum bounding box size of 36 pixels for all images and frames. The evaluation harness can properly run all protocol files and generate and plot the required metrics.

Valid participation in the IJB-A detection benchmark requires reporting mean and standard deviation (measured across all splits) for the following: (i) the accuracy metrics, and (ii) detection duration of all imagery in a split. These are the two primary factors for operational deployment [4]. Research must also confirm that the protocol was precisely followed, and that a minimum bounding box size of 36 pixels was used for all images and video frames.

4.2. Baseline

Baseline results are provided in this section for the following detectors: (i) OpenCV's Viola Jones Haar cascade [18] (VJ), (ii) Dlib's HOG-based detector[7], and (iii) a government off the shelf detector (GOTS). Both VJ and Dlib are available in the open source. The GOTS detector was the top performing detector in a recent benchmark [4], where it was shown to achieve results similar to the top published methods on FDDB [7]. For the Viola Jones detector, the pre-trained "alt2" model was used, as it achieved the best detection accuracy over all models provided in OpenCV. Providing this range of detectors allows researchers to understand how their method performers against different tiers of availability. The mean and s.d. for detection accuracies can be found in Table 4. Plots and other metrics will be provided on the database website.

5. Summary

This paper has introduced the IARPA Janus Benchmark A (IJB-A) dataset. The IJB-A is a joint face detection and face recognition dataset. The dataset consists of face images and videos that were collected "in the wild". A key distinction between this dataset and previous datasets is that all faces have been manually localized and have not been filtered by a commodity face detector. The implication of this approach is an unprecedented amount of variation in pose, occlusion and illumination in the IJB-A dataset.

The IJB-A dataset is motivated by a need to push the state of the art in unconstrained face recognition. As such, it allows improvements in the face recognition tasks to proceed without being blocked by parallel research in face detection. The recognition protocol provides a distinction for whether or not training was performed on the gallery. While training on a gallery introduces potential computation bottlenecks in image enrollment, it allows for a paradigm to be explored that is similar to familiar human recognition of "familiar" faces.

Finally, expansion of the benchmark dataset will be released over time to facilitate further unconstrained face analysis. Such expansions will include, at the least, a denser set of fiducial landmarks, attribute labels, additional subjects, and protocols for open set face recognition. Researchers can subscribe the (annonymized) mailing list for information on such future releases.

Table 4. Face detection accuracies on the proposed IJB-A dataset at specified operating points. Shown are the true detect rates (TDR) at false detect rate (FDR) per image of 0.1 (one false detect every 10 images) and 0.01 (one false detect every 100 images).

Detector	FDR = 0.1	FDR = 0.01
VJ [18]	0.370 ± 0.011	0.173 ± 0.012
Dlib	0.497 ± 0.015	0.165 ± 0.014
GOTS	0.765 ± 0.008	0.259 ± 0.058

[7] http://dlib.net

References

[1] *IARPA Janus Broad Agency Anouncement, IARPA-BAA-13-07*. 1

[2] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security*, 2014. 3

[3] J. R. Beveridge et al. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE Biometrics: Theory, Applications, and Systems*, 2013. 4

[4] J. Cheney, B. Klein, A. K. Jain, and B. F. Klare. Unconstrained face detection: State of the art baseline and challenges. In *IAPR Int. Conference on Biometrics*, 2014. 2, 7, 8

[5] P. Grother and M. Ngan. Face recognition vendor test (FRVT): Performance of face identification algorithms. In *NIST Interagency Report 8009*, 2014. 2, 5, 6

[6] 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, October 2007. 1, 4, 6

[7] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010. 7, 8

[8] B. Klein, K. Allen, A. Jain, and B. Klare. Limiting factors in unconstrained face recognition and detection. In *IEEE Biometrics: Theory, Applications, and Systems (under review)*, 2015. 4

[9] J. C. Klontz and A. Jain. A case study of automated face recognition: The boston marathon bombing suspects. In *IEEE Computer*, November 2013. 1

[10] J. C. Klontz, B. F. Klare, S. Klum, A. K. Jain, and M. J. Burge. Open source biometric recognition. In *IEEE Biometrics: Theory, Applications, and Systems (under review)*, 2013. 7

[11] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009. 1, 4

[12] S. Liao, Z. Lei, D. Yi, and S. Z. Li. A benchmark study of large-scale unconstrained face recognition. In *International Joint Conference on Biometrics (IJCB)*, 2014. 2

[13] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014. 2

[14] V. Natu and A. J. OToole. The neural processing of familiar and unfamiliar faces: A review and synopsis. *British Journal of Psychology*, 102(4):726–747, 2011. 2

[15] A. J. O'Toole, X. An, J. Dunlop, V. Natu, and P. J. Phillips. Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception (TAP)*, 9(4):16, 2012. 1

[16] E. Taborsky, K. Allen, A. Blanton, A. K. Jain, and B. F. Klare. Annotating unconstrained face imagery: A scalable approach. In *IAPR Int. Conference on Biometrics*, 2014. 4

[17] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Computer Vision and Pattern Recognition*, pages 1701–1708. IEEE, 2014. 1, 2

[18] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 2, 8

[19] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Computer Vision and Pattern Recognition*, pages 529–534. IEEE, 2011. 1, 4, 6