To Appear: The 8th IAPR International Conference on Biometrics (ICB), Phuket, Thailand, May 19 - 22, 2015.

Unconstrained Face Detection: State of the Art Baseline and Challenges

Jordan Cheney* Ben Klein* Anil K. Jain† Brendan F. Klare*

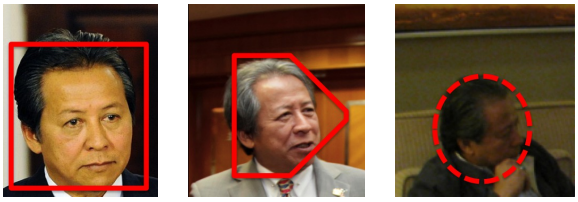
Abstract

A large scale study of the accuracy and efficiency of face detection algorithms on unconstrained face imagery is presented. Nine different face detection algorithms are studied, which are acquired through either government rights, open source, or commercial licensing. The primary data set utilized for analysis is the IAPRA Janus Benchmark A (IJB-A), a recently released unconstrained face detection and recognition dataset which, at the time of this study, contained 67,183 manually localized faces in 5,712 images and 20,408 video frames. The goal of the study is to determine the state of the art in face detection with respect to unconstrained imagery which is motivated by the saturation of recognition accuracies on seminal unconstrained face recognition datasets which are filtered to only contain faces detectable by a commodity face detection algorithm. The most notable finding from this study is that top performing detectors still fail to detect the vast majority of faces with extreme pose, partial occlusion, and/or poor illumination. In total, over 20% of faces fail to be detected by all nine detectors studied. The speed of the detectors was generally correlated with accuracy: faster detectors were less accurate than their slower counterparts. Finally, key considerations and guidance is provided for performing face detection evaluations. All software using these methods to conduct the evaluations and plot the accuracies are made available in the open source.

1. Introduction

The release of unconstrained face recognition datasets such as the Labelled Faces in the Wild (LFW) [7] and YouTube Faces [17] datasets represented a significant challenge to face recognition algorithms at the time of their release. However, recognition accuracy has recently begun to saturate on these datasets [15, 5, 13]. At the same time, the challenge of such unconstrained face datasets is diminished by the fact that all faces are detectable by a commodity face detection algorithm.

Recently, the IAPRA Janus Benchmark A (IJB-A) was released, which consists of labelled face images that were all manually localized [9]. In addition to representing a significantly more difficult recognition challenge, many of the faces in this dataset cannot even be detected consistently. As the early generation of unconstrained face recognition datasets continue to saturate, and more unconstrained face datasets emerge (i.e., images with manually detected faces), face detection may become a bottleneck for face recognition. As such, it is critical to understand the state of the art in face detection accuracy of stable, "off the shelf" detectors.



(a) (b) (c)

Figure 1. Examples of different ranges of difficulty for automated face detection and recognition. (a) Faces with limited pose variations. (b) Faces with a larger range of pose variation. (c) Full pose variation.

While prior to the IJB-A dataset no manually localized "media in the wild" face recognition dataset existed, several manually localized unconstrained face detection databases existed, such as FDDB [8] and AFLW [10]. Using such datasets, a recent study demonstrated that significant challenges remain in unconstrained face detection [12]. Additionally, it was shown that detection protocols and evaluations metrics can greatly influence measured detection accuracy. Finally, a comparison of two different algorithmic approaches, rigid templates [16] and deformable parts-based models [4] was conducted, with the finding that both methods can yield top accuracy across multiple benchmarks. The work in this paper compliments previous findings [12], and proceeds in orthogonal directions by: (i) measuring the accuracy of the off the shelf detectors on a much larger scale of test imagery, (ii) introducing additional evaluation considerations when benchmarking face detection algorithms, and (iii) offering several additional factors that still confound face detection algorithms.

The contributions of this study are as follows: (i) the largest scale, off the shelf face detection evaluation in terms of both number of detectors studied, nine, and dataset size, (ii) an examination of both accuracy and speed, and (iii)

*J. Cheney, B. Klein, and B. Klare are with Noblis, Falls Church, VA, U.S.A.

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

Name	Abbreviation	Domain
Commercial Detector I	COTS-I	Proprietary
Government Detector I	GOTS-I	Government
Government Detector II	GOTS-II	Government
Government Detector III	GOTS-III	Government
Academic Detector I	MSU	Academic
OpenCV	VJ	Open Source
Dlib	DLIB	Open Source
PittPatt ver. 4	PP4	Government
PittPatt ver. 5	PP5	Government

Table 1. Face detection algorithms evaluated in this study.

identifying factors that still confound face detection accuracy. All code used to evaluate and plot the detection results are provided in the open source[1].

The remainder of the paper is organized as follows. Section 2 provides information regarding the evaluation methodology, datasets, and face detectors. Section 3 contains detection results on the IJB-A dataset, and Section 4 contains results on the FDDB datatset. The computational costs for each detector are presented in Section 5. Broad analysis and observations are offered in Secton 6. Finally, we summarize the observations of this study in Section 7.

2. Methodology

2.1. Detectors

Nine face detection algorithms were evaluated in this study. These detectors were developed across commercial, academic, and open source organizations. A listing of all detectors can be found in Table 1. The name of the commercial off the shelf (COTS) detector has been withheld due to licensing agreements with the vendor. However, this detector is from a face recognition system that was in the upper echelon of the algorithms evaluated in the recent NIST FRVT study[6]. One important consideration for COTS detectors is that they are generally tuned to detect faces that are of recognition quality. Similarly, the names of the three government off the shelf (GOTS) detectors have been withheld; they were selected based on their previously reported performance. Two open source face detection algorithms are tested: OpenCV's Viola Jones-based detector [16], and the detector provided in the Dlib library (`http://dlib.net/`). It is important to note that for OpenCV the pre-packaged `alt2` model was used instead of the `default` model, as it was found to be the most accurate pre-packaged model (see Figure 7(b) for such comparison). For the OpenCV detector, the minNeighbors parameter was tuned and set to 5.

All of studied detectors were used as pre-trained detectors. Because the IJB-A dataset was not released prior to this study, it operates as sequestered data with which to evaluate these detectors. The technical approaches for each of the studied detector are not always known as certain methods are proprietary. However, the following information is known: The OpenCV detector represents an image using Haar features and classifies as region using cascaded decision stump classifiers. The DLib detector uses a histogram of oriented gradients (HOG) represetnation in conjunction with a SVM classifier [3]. The MSU detector uses a normalized pixel difference (NPD) representation in conjunction with a cascaded regression tree classifier [11]. The other detectors are proprietary and specific implemenation details are not available.

[1] `openbiometrics.org`

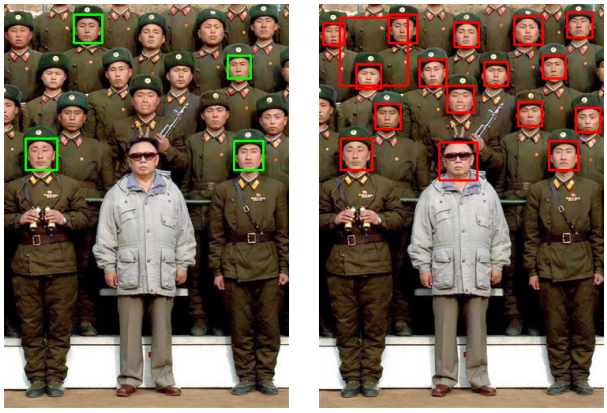
Dataset	Image Area (px)	Face Size (px)	Faces / Image
IJB-A Images	$4.04e6 \pm 5.07e6$	249.4 ± 277.4	2.9 ± 3.2
IJB-A Frames	$4.45e5 \pm 3.94e5$	120.3 ± 88.5	1.7 ± 1.7
FDDB	$1.47e5 \pm 1.85e4$	94.1 ± 49.3	1.8 ± 1.5

Table 2. Statistics of the three different image and video sets studied.

2.2. Datasets

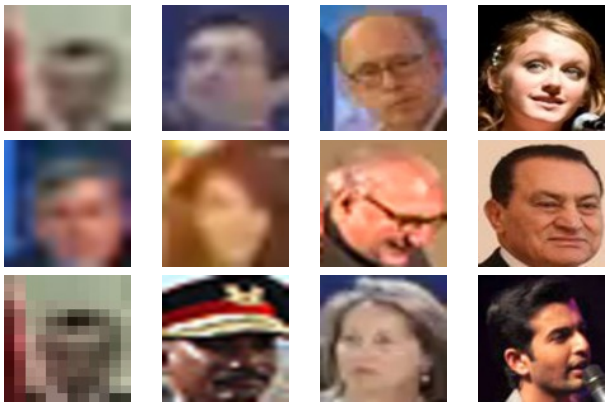
IJB-A Dataset The IJB-A is a recently released "media in the wild" dataset that consists of 67,183 manually localized faces across 5,712 images and 20,408 video I-frames (sampled from 2,805 video clips) [9]. The IJB-A data was manually collected from open source image and video collections. Table 2 contains relevant statistics on the IJB-A subsets and FDDB for reference. The primary purpose of the dataset is to evaluate the accuracy of face recognition (not detection) algorithms. As such, all images and videos in the dataset are labelled with the subject identity of at least one person, and the labelled subjects have manually annotated landmarks for the center of the two eyes and the base of the nose. However, a key feature of the IJB-A dataset is that the location of all faces were manually localized and annotated in order to develop face recognition algorithms that are robust to the full variations in pose, illumination and occlusion. As such, the IJB-A imagery is the first known dataset that can be used for both face detection and recognition. This report focuses solely on face detection.

FDDB In addition to the IJB-A dataset, detection results on the FDDB (Face Detection Data Set and Benchmark) [8] are presented. The FDDB database consists of 2,845 images containing a total of 5,171 faces. The images were collected from news articles on the Yahoo website, and all faces were manually localized for ground truth. The FDDB dataset has been selected in this study because it is the most widely used benchmark for unconstrained face detection, which allows the off the shelf detectors in this study to be compared against the academic self reported algorithms. Further, a smaller scale evaluation of available face detectors was recently conducted which evaluated some of the same face detectors in this study on the FDDB benchmark [1]. Most notable from these distributions is that the image and face



(a) (b)

Figure 2. Output of the VJ detector at specific false positive rates. (a) False positive rate of roughly 0.01 or 1 false positive per every 100 images. (b) False positive rate of roughly 0.1 or 1 false positive per every 10 images. As is shown in the figures a higher false positive rate results in far more detections, but at the expense of introducing false positives.



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

Figure 3. Manually annotated faces from the IJB-A dataset with bounding boxes that are (a) 12 pixels, (b) 20 pixels, (c) 36 pixels, and (d) 72 pixels along their smallest side (height or width). The results presented on the IJB-A dataset are on faces that are 36 pixels and above, and results on the FDDB dataset are on faces 20 pixels and above.

sizes are significantly smaller in FDDB than in IJB-A.

2.3. Evaluation Harness

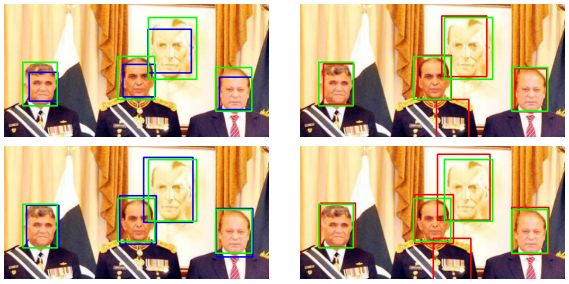
The evaluation and plotting functionality embedded within the OpenBR project was used to analyze the face detection accuracies reported in this study. Detector accuracies are reported as receiver operating characteristics (ROC). True accept rates for ROC curves are plotted as a function of false accepts per image, which differs from the FDDB benchmark which reports the total false accept rate. For example, at 0.1 false accepts / image, an average of one false accept occurs per every ten images. In this evaluation, two false accept rates per image, 0.1 and 0.01, were selected as being particularly relevant to operational applications. Figure 2 illustrates the difference between these two operating thresholds. The reason for presenting false accepts per image, instead of total false accepts, is that false accepts per image is invariant to the number of images in a given dataset. Thus, cross dataset comparisons are better facilitated using this reporting metric, in addition to having a more meaningful interpretation for readers.

2.4. Minimum Bounding Box Size

A key, and often overlooked, consideration when comparing face detection algorithms is the default minimum bounding box size used by a given detector. If two algorithms being compared use different minimum bounding box sizes, then the detection results may not properly convey which detector is more accurate. For example, if the minimum bounding box size is smaller than the size of ground truth face sizes, then searching for faces at such smaller locations can only increase the false positive count, and not the true positive count. Conversely, if the minimum bounding box size is larger than the size of the ground truth face sizes, then the true positive count will be restricted from including such smaller faces in the data. Similarly, false positive rates will vary if the minimum bounding box size is set to a larger value.

Two different minimum box size sizes where used with the IJB-A and FDDB datasets. For FDDB, a minimum bounding box size of 20 was used, as this is the protocol provided by this dataset. For the IJB-A dataset, faces have been annotated as small as 12 pixels; however, the ground truth was carefully filtered (see next paragraph) to only contain 36 pixels or larger. The choice of 36 pixels is motivated by our application of interest (face detection in conjunction with face recognition). As demonstrated in NIST's FRVT evaluation released in 2014 [14], the accuracy of most face recognition algorithms precipitously decreases at an interpupilary distance of 18 pixels, which roughly corresponds to a bounding box size of 36 pixels. Thus, a minimum bounding box size of 36 was set for all detectors when operating on IJB-A. Examples of varying appearance of different sized faces can be found in Figure 3.

A naive implementation of this filter could just remove bounding boxes from the ground truth and detector output if they are smaller than the threshold. This method however, unfairly benefits algorithms that output larger bounding boxes. As an example, consider a head-cropped (encompasses the entire head and not just the facial landmarks) ground truth bounding box of size 36 and two detectors, one which returns a crop around the head and one which returns a crop around the facial landmarks. Using the naive method, the facial crop would be filtered out if the threshold was set to 36 pixels, while the head crop would remain, crediting an incorrect false positive to the detector that output smaller boxes. To avoid this, we introduce a second threshold, introduced in [12], for the predicted bounding boxes. This



(a) (b)

Figure 4. When evaluating pre-trained face detection algorithms, different bounding box sizes may inappropriately influence measured accuracy. As such, bounding box sizes need to be normalized on a per detector basis. The top row shows raw bounding box results (a) VJ (in blue) and (b) PP4 (in red). In each image the green box indicates the manually annotated ground truth. The bottom row shows the boxes after applying the normalization which fixes biases in different box sizes (a) and does not alter well aligned boxes (b).

new threshold, β, relates to the original threshold, α, such that

$$\beta = \sqrt{0.5 * \alpha^2} \tag{1}$$

The purpose of β is to keep all predicted bounding boxes that could overlap the ground truth by 50%, the minimum overlap required for a detection to be considered a true positive. By introducing this second threshold, we are able to more intelligently crop the output of each detector and avoid the situation described above.

Finally, it is important to note that by *default*, the minimum bounding box size was different for each detection algorithm. In order to properly compare these detectors, these values had to be manually set. While some detectors expose this parameter as minimum bounding box size, others specify this value as a percentage of image width: thus, care needed to taken with all nine detectors to ensure the minimum bounding box size was properly set.

2.5. Bounding Box Normalization

An output from a given detector is considered to match a ground truth face location if the two bounding boxes overlap by at least 50%. When analyzing pre-trained face detectors, variations in bounding box sizes across detectors is another factor that needs to be considered in order to properly compare multiple detectors. For example, the top row of Figure 4 shows the bounding boxes calculated by two detectors (VJ and PP4), as well as the ground truth locations for the face images. While the bounding box dimensions of certain detectors are closely aligned with the ground truth (in this example, PP4), others are consistently of different shape and localization (in this example, VJ). Without considering these differences, a detector that is not closely aligned with the ground truth dimensions will be reported as being less accurate when instead it could be an issue of the different bounding box formats causing certain face detections to have less overlap with the ground truth.

In order to properly compare the detectors, bounding box normalization was applied to each predicted bounding box. Bounding boxes in this study are defined in terms of the top left corner, given as an x and y coordinate, and a width and height. The goal of the normalization algorithm is to determine four coefficients, dX, dY, $dWidth$, and $dHeight$, for each detector. These values correspond to fixed scale and translation transformations that can be applied to each bounding box output such that each box is scaled to more closely match the ground truth.

The following algorithm is applied to the output of each detector individually at evaluation time to properly normalize faces. First, two sets of rectangles R'_T and R'_P are generated, where R'_T corresponds to ground truth faces, and R'_P corresponds to the predicated output by the detector. From these two sets, two new sets, R_T and R_P, are generated which contain the boxes that have a minimum overlap of at least $\tau, 0 < \tau \leq 1$. We chose $\tau = 0.3$ based on empirical examination. The overlap of two rectangles was measured by dividing the area of intersection between the two rectangles by the area of the union of the two rectangles. The i-th ground truth and predicted box are denoted as $R_T(i)$ and $R_P(i)$, respectively, where $1 \leq i \leq n$. Next, we compute the n-dimensional vector of differences x_l, x_r, x_t, and x_b, which correspond to the width normalized difference between the left, right, top, and bottom dimensions of the ground truth and predicted rectangles. That is:

$$x_l(i) = \frac{R_T(i).l - R_P(i).l}{R_P(i).w} \tag{2}$$

$$x_r(i) = \frac{R_T(i).r - R_P(i).r}{R_P(i).w} \tag{3}$$

$$x_t(i) = \frac{R_T(i).t - R_P(i).t}{R_P(i).w} \tag{4}$$

$$x_b(i) = \frac{R_T(i).b - R_P(i).b}{R_P(i).w} \tag{5}$$

where $*.w$ denotes rectangular width, $*.l$ denotes the left boundary coordinate, $*.r$ the right, $*.t$ the top, and $*.b$ the bottom. Finally, we normalize each predicated box location into the normalized set $R_N(i)$ as:

$$R_N(i).l = R_P(i).l + \frac{x_l(i) \cdot R_P(i).w}{n} \tag{6}$$

$$R_N(i).r = R_P(i).r + \frac{x_r(i) \cdot R_P(i).w}{n} \tag{7}$$

$$R_N(i).t = R_P(i).t + \frac{x_t(i) \cdot R_P(i).w}{n} \tag{8}$$

$$R_N(i).b = R_P(i).b + \frac{x_b(i) \cdot R_P(i).w}{n} \tag{9}$$

The bottom row of Figure 4 shows these boxes after applying our normalization algorithm. As demonstrated, the normalization causes successful detections to more closely align with the ground truth.

Detector	FAR = 0.01	FAR = 0.1
PP4	0.28	0.78
PP5	0.25	0.75
COTS-I	0.21	0.48
GOTS-I	0.03	0.38
GOTS-II	0.19	0.52
GOTS-III	0.14	0.37
VJ	0.17	0.38
DLIB	0.18	0.52
MSU	0.31	0.50

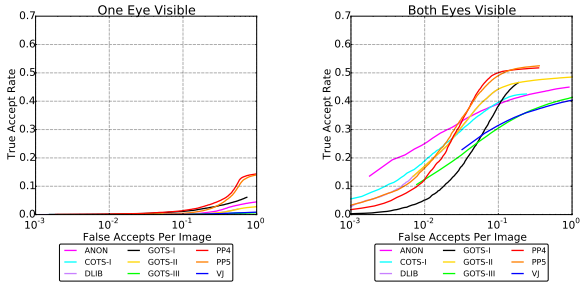
Table 3. Detector performance on the entire IJB-A Dataset. Listed are true accept rates (TAR) at false accept rates (FAR) of 0.01 (corresponding to one falsely accepted face per 100 images) and 0.1 (corresponding to one false accepted face per 10 images).

3. IJB-A Dataset Results

This section provides the detection accuracies results for the nine commercial, government, and open source detectors that were evaluated on the IJB-A dataset. The detection performance of all detectors on IJB-A are contained in Figure 5; Table 3 contains the corresponding accuracies at key operating thresholds for the entire image corpus. Because the IJB-A dataset consists of two distinct subsets, still images and video frames, Figure 5 contains plots for both the entire dataset, as well as each subset individually. All results shown are based on a minimum bounding box size of 36 pixels. Thus, each detector was set to not search for smaller faces, and any face smaller than 36 pixels was removed from the ground truth. Finally, the outputs from all detectors were normalized using the normalization algorithm described in Section 2.5.

Key observations from these results are as follows. The two PittPatt detectors (PP4 and PP5) exhibit the best detection performance at high FARs (> 0.1) by a wide margin. Additionally, PittPatt 4 performs better than PittPatt 5 at all FARs (this will be discussed in more detail later). The open source VJ detector is one of the worst detectors at all operating points. The DLIB detector, which is also open source, shows much better performance. GOTS-I is the worst detector by a wide margin at low FARs but shows rapid improvement as FAR increases. At higher FAR values (> 0.1) it is a top performing detector. The MSU detector is the best at low FARs. At higher FARs its performance is lower relative to other detectors.

At least one face in each image and video frame in the IJB-A dataset has landmarks denoting the center of the eye sockets and base of the nose annotated using crowdsourced labor from Amazon Mechanical Turk. Such information was collected for the labelled person of interest in each image as the dataset is also used for evaluating face recognition. In the context of face detection, this information can be used to coarsely categorize the pose of certain faces. Specifically, we consider two cases: (i) both eyes are visible, and



(a) (b)

Figure 6. Detector performance on (a) subjects with one eye visible and (b) subjects with both eyes visible

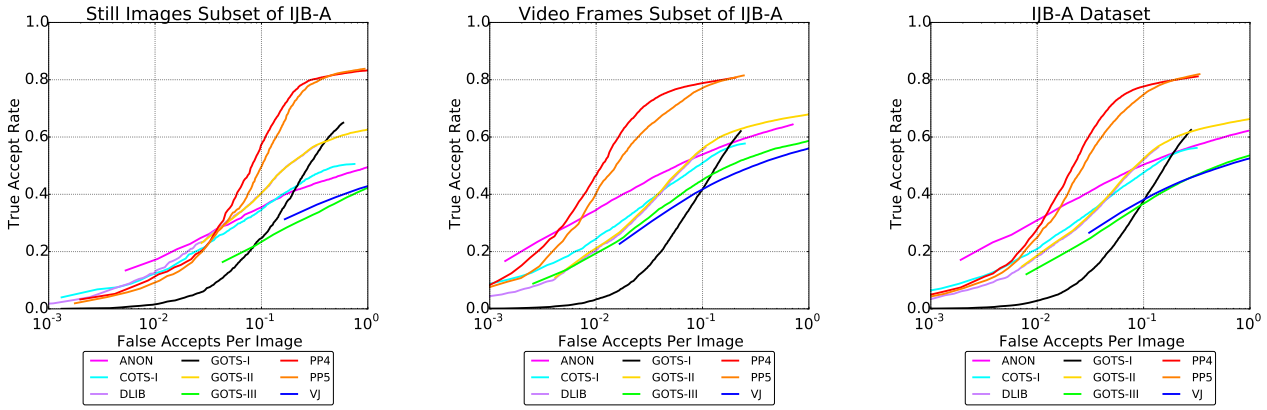
(ii) only one eye is visible. When both eyes are visible, this roughly corresponds to yaw ranges of $-45°$ and $45°$. Faces without landmark annotations (i.e. faces not belonging to the identity labelled subject in the image) were marked with an *IGNORE* flag. This means that successful detections of these faces did not count as false positives for the detectors, and unsuccessful detections did not count as true negatives.

Figure 6 shows that pose still has a massive effect on the performance of all face detectors. Further, the PittPatt detectors were top performers on both more frontal and more extreme pose variations. GOTS-I was the best non-PittPatt detector on the extreme pose bins at all FARs. On frontal pose bins it was the worst performer at lower FARs, but exhibited strong relative improvement as FAR increased. This is the same trend that the detector showed on the full IJB-A dataset. MSU performed in the top half of detectors on extreme pose bins and was the best detector at low FARs on frontal pose bins. Again, this mirrors the behavior on full IJB-A. VJ performs in the bottom half of detectors on both pose bins, and again performs worse than an open source alternative (DLIB) at all FARs.

4. FDDB Dataset Results

This section contains the detection performance for the nine detectors on the FDDB dataset. Because many of the evaluated detectors do not disclose which datasets they were trained on, it is important to consider that certain detectors may have been trained on FDDB (which would increase their reported accuracy). All detectors were evaluated with a minimum bounding box size of 20 pixels, which is the smallest ground truth bounding box reported in FDDB. The output from each detector was normalized using the normalization algorithm described in Section 2.5.

Figure 7(a) shows the ROC curves for the evaluated detectors on the FDDB dataset. The accuracies observed using the `alt2` VJ model and the accuracies reported on the FDDB benchmark (most likely using the `lbp` model) are quite different. As such, Figure 7(b) shows the ROC curves for the detection performance of different OpenCV models compared to the reported results from FDDB. The difference in accuracy is mostly attributed to the different model



(a) (b) (c)

Figure 5. Detector performance on IJB-A dataset. Plots shows true accept rate (TAR) vs. false accepts per image for (a) still images, (b) video frames, and (c) images and videos combined.

file used. Figure 7(c) shows the ROC curves for the top six self-reported results in the published literature compared against the top six evaluated detectors on the FDDB dataset in this study. At the time of publication, the best performing algorithm on FDDB was the joint cascade approach [2]. One important note is that the self-reported detectors were evaluated outside of the evaluation harness used in this paper. The methodology for that evaluation is specified in the FDDB evaluation standard and differs from the evaluation methodology of this paper.

The most notable observation is that top off the shelf detectors perform at roughly the same level as academic self reported algorithms. This may be partially due to certain algorithms having been trained on FDDB as these are all pre-trained detectors.

Key observations from these results are as follows. The PittPatt detectors were the top performers on FDDB, with PP4 once again outperforming PP5. COTS-I is the third highest performing detector, this is a higher relative performance than was shown on IJB-A. GOTS-I performs much better at low FARs on FDDB than on IJB-A and is a top performing detector at all FARs. VJ is the lowest performing detector on FDDB. The performance shown however is much higher then was self reported on the FDDB website. This is due to a different model file being used, see Figure 7(b). MSU exhibits lower relative performance on FDDB than on IJB-A.

5. Runtime

In any operational deployments, the face detection accuracy is not the only consideration for overall performance. Other factors, such as memory usage, computational efficiency, and (especially) detection time are also critical. Applications such as face detection in cameras on mobile phones often prioritize speed over detection accuracy, while security applications might find value in the detection of every face, at the expense of real-time operation. For these reasons, the following section will provide computational benchmarks for each detector.

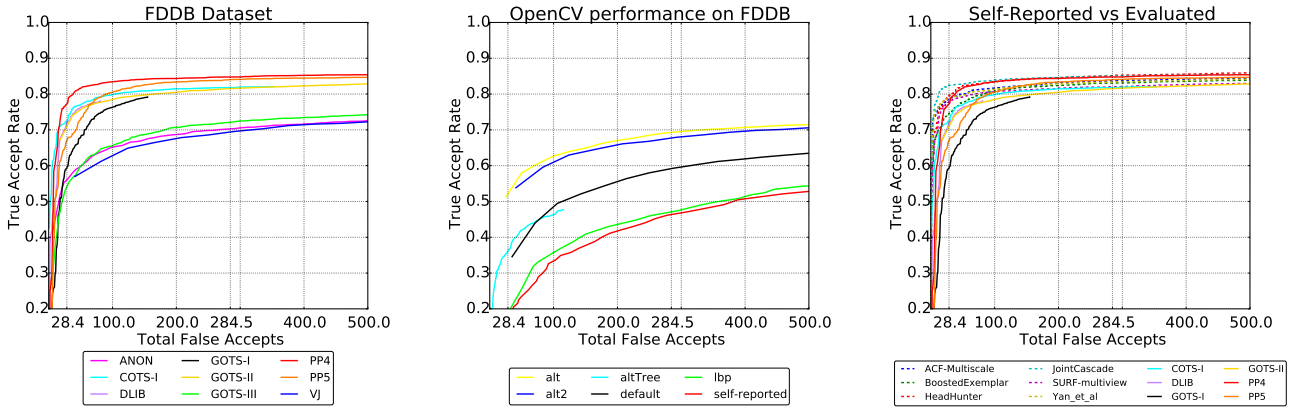
All detectors were run in a single threaded environment. Time measurements were performed on randomly selected splits from both the still images and video frames subsets of the IJB-A dataset, where each split contained 200 images and the splits had non-overlapping imagery. Non-detection overhead, such as opening an image or writing an image to disk, was handled outside of the profiling and did not effect the measurements. Results are reported as average face detection time per image.

Figure 8(a) shows the average time per image for all random splits of the video frames subset of IJB-A. Figure 8(b) shows the detection performance at a false positive rate of 1 false positive per 10 images measured against the computational performance.

The tradeoff between detection performance and the computational performance can be summarized as follows. The detectors fell into two tiers of speed, with the slower tier having the most accurate detectors. PP4 was the most accurate detector and was in the middle of the slower tier. PP5 was the second most accurate detector and was the fastest detector in the slow tier. This speed increase could explain the accuracy difference between PP4 and PP5. MSU was the fastest overall detector and also the most accurate detector in the faster tier. GOTS-I was one of the worst performers and was the slowest detector. GOTS-II and DLIB performed very similarly. They were average in both performance and speed in the slower tier. VJ was the second fastest detector but also one of the worst performers.

6. Analysis and Observations

Pose was observered as the most important factor in detector performance. This was evident by the results shown in Figure 6. However, additional qualitative analysis of failure cases further demonstrates this trend. For example, Figure 9 show faces that were detected by all detectors, and



(a) (b) (c)

Figure 7. (a) Detector performance on the FDDB dataset. Tick marks at 28.45 and 284.5 correspond to false accept rates of 0.01 and 0.1, respectively. (b) Performance of different OpenCV models with self-reported results on FDDB dataset. (c) Comparison of top evaluated results and top self-reported results on FDDB dataset. Off the shelf detectors exhibit more false accepts than self reported algorithms, which may be due to self reported methods ignoring smaller bounding box sizes.

face detected by none of the detectors. While these were carefully selected to demonstrate a range of variates that cause all detectors to fail, the dominance of pose was clear when inspecting all of the 13,872 faces (or 20.65%) that failed to be found by all detectors. As shown in Figure 9, other predominant factors that impact face detection accuracy are partial occlusion, and poor illumination.

Computational speed is a strong indicator of detector performance. The most accurate detectors are in a slower tier (in general) then the less accurate detectors.

Image resolution plays a very large factor in detector performance and computational speed. High resolution images will generate a significantly higher false positive rate. This is shown in the general shift to the right for detector performance between Figure 5(a) and Figure 5(b). Additionally, higher resolution imagery will slow down the detector as there is far more area to search. As such, understanding resolution and the average size of faces in the target imagery can help lower computation time and raise detection performance by adjusting minimum bounding box size.

The fusion of multiple was studied in this research, however the top detector (PP4) did not benefit from being fused with any other detector.

7. Conclusions

An evaluation of nine different face detection algorithms was performed. The detectors were sourced from industry, government, open source and academia. Both the accuracy and efficiency of these algorithms were measured on two datasets, one containing over 67,183 faces (the IJB-A dataset) and the other used as a common benchmark in academic research. From this evaluation it is clear that while significant progress has been made in face detection, several challenges still remain. Most notable of these challenges is robustness to facial pose.

With over 20% of the faces in IJB-A failing to be detected by any one of the nine detectors, it is clear that more novel approaches to face detection are needed to advance state of the art in face recognition. Such approaches will need to range across both face representations and learning algorithms. We may potentially have to use several face detectors trained on sets of faces containing more homogeneous poses.

Acknowledgement

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under contract number 2014-14071600010. 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 purpose notwithstanding any copyright annotation thereon.

References

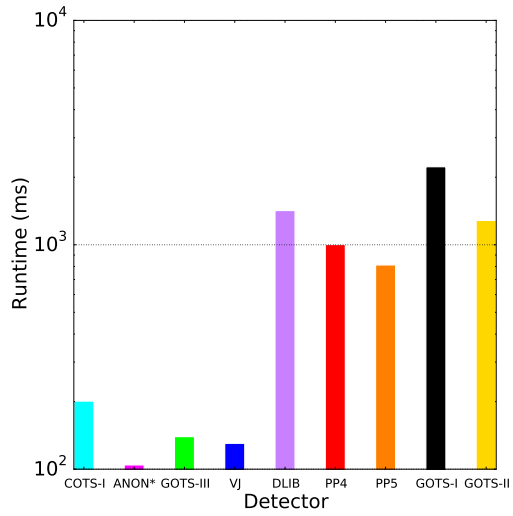
[1] K. Brady. Face detection technology evaluation. In *Project Report (DOMEXP-1)*, 2014.

[2] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *Proc. of ECCV*, pages 109–122, 2014.

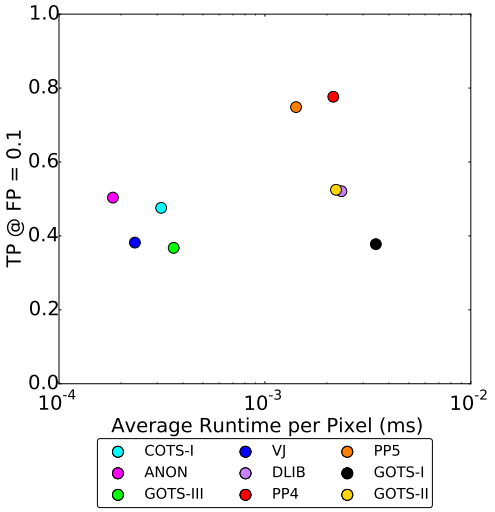
[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, 2005.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE PAMI*, 32(9):1627–1645, 2010.

[5] G. Goswami, R. Bhardwaj, R. Singh, M. Vatsa, and I. IIIT-Delhi. Mdlface: Memorability augmented deep learning for video face recognition. In *IJCB*, pages 1–7, 2014.



(a)



(b)

Figure 8. (a) Average runtime speed for each detector. (b) Comparison of detection accuracy performance versus runtime *per pixel*, where detection accuracy is reported as true positive rate at false positive rate per image of 0.1.

(a)

(b)

Figure 9. (a) shows examples of faces that were not detected by any face detector. (b) shows examples of faces that were detected by all of the face detectors.

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

[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. In *Tech Report 07-49, Univ. of Massachusetts*, 2007.

[8] 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.

[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 *Proc. of CVPR (to appear)*, 2015.

[10] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[11] S. Liao, A. K. Jain, and S. Z. Li. Unconstrained face detection. In *MSU Technical Report, MSU-CSE-12-15*, 2012.

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

[13] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In *Proc. of CVPR*, 2014.

[14] G. W. Quinn and P. J. Grother. Performance of face recognition algorithms on compressed images, nist interagency report 7830. NIST, 2011.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. of CVPR*, 2014.

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

[17] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. of CVPR*, 2011.