2017 IEEE 12th International Conference on Automatic Face & Gesture Recognition

A Cross Benchmark Assessment of A Deep Convolutional Neural Network for Face Recognition

P. Jonathon Phillips
National Institute of Standards and Technology
Gaithersburg, MD 20899 USA

***Abstract*— Deep convolutional neural networks (DCNN) based algorithm methods have swept face-recognition. DCNN-based algorithms have shown significant improvements in accuracy on the Labeled Faces in the Wild (LFW) and the YouTube[1] Video face-recognition benchmarks. These two benchmarks consist of images and videos of celebrities downloaded from the World Wide Web. Since 2004, the National Institute of Standards and Technology (NIST) has established a series of face-recognition benchmarks that span a range of scenarios and difficulties. The scenarios range from comparing frontal faces taken in studio lighting to comparing faces acquired with cell phone cameras taken outdoors. The VGG-face algorithm [7] was ran on eight NIST face-recognition benchmarks. The Vision Geometry Group (VGG)-face algorithm excelled on the most difficult benchmarks; existing algorithms excelled the benchmarks with higher quality images. This finding is consistent with the design of the algorithms. The VGG-face algorithm was designed to recognize faces in variable illumination; the existing algorithms were designed to operate on face-images taken in controlled illuminations. To accurately characterize the performance of face recognition algorithms, we recommend that performance is reported on multiple benchmarks.**

I. Introduction

Deep convolution neural networks (DCNN) have reported extremely high accuracy rates on the Labeled Faces in the Wild (LFW) and YouTube Video face recognition benchmarks [3], [15]. The LFW benchmark is typical of the popular benchmarks released since 2006. The images are downloaded from the World Wide Web and consist of pictures of celebrities taken in ambient lighting. The LFW benchmark opened the internet as a source of images and allowed researchers to address the problem of recognizing faces that can be found on the Web. The accuracy on the LFW has improved from 0.78 in 2008 to near perfection today. The increase in performance is due to advances in face recognition techniques that were motivated by improving accuracy on the LFW. The most accurate algorithms were based on DCNNs.

Since 2004, the National Institute of Standards and Technology (NIST) has established a series of face-recognition benchmarks that span a range of scenarios and difficulties. The scenarios range from comparing frontal faces taken in studio lighting to comparing faces acquired with cell phone cameras taken outdoors. These benchmarks measure the performance on a wider range of conditions than the LFW and YouTube faces datasets. We report performance of the Vision Geometry Group (VGG)-face algorithm [7] on eight NIST benchmarks. The VGG-face algorithm is a high performing algorithm on the LFW benchmark. The eight benchmarks span a range of scenarios from face images acquired in a studio environment to face images taken with digital point and shoot cameras. Our analysis for the first time will assess a DCNN-based algorithm accuracy on a range of conditions, compare the results to established accuracy on each of the benchmarks, and compare the DCNN algorithm accuracy to human accuracy. Based on our analysis, we recommend changes in assessing the accuracy of face recognition algorithms.

[1]The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST.

II. Overview of Benchmarks

The images in the eight NIST benchmarks were all collected at the University of Notre Dame between 2002 and 2011 [9]. For seven of the benchmarks, the images were acquired with a digital single lens reflex camera. The images were taken in studio lighting and in ambient lighting in hallways and outside. Humans would consider the images high quality. The images in one benchmark were acquired with digital point and shoot cameras. The overall demographic composition over the eight benchmarks was 59% male and 41% female; 71% Caucasian and 10% East Asian; and 92% were 18 to 29 years old. The demographics varied slightly by benchmark with precise numbers provided in the references.

In the benchmarks, performance was only measured for frontal face images. There are two performance metrics for assessing progress on the benchmarks. The first is verification rate (VR) at a false accept rate (FAR) of 1 in 1000. This statistic was selected to compare algorithms. The relatively low false accept rate was selected because in real-world applications, there is a desire to minimize the number of false accepts. In addition to this statistic, performance is plotted on a receiver operating characteristic (ROC). The second figure of merit is the area under the ROC (AUC). AUC is reported when a benchmark is designed to compare human and algorithm performance.

III. VGG-face algorithm

VGG-face is a complete face recognition system that includes a deep convolution neural network for recognition and a deformable parts model (DPM) for face detection

978-1-5090-4023-0 2017

DOI 10.1109/FG.2017.89

and localization [2], [4]. The experiments in this paper used the Matlab implementation available from the VGG website. The CNN distributed was pretrained weights. The basic architecture of the algorithm consists of 37 layers, and the layers include convolutions, relu normalization, pooling, fully connected layers, and softmax. The accuracy measure on LFW and YouTube faces dataset is 1.0-ERR, where ERR is equal error rate. The accuracy of the VGG-face on the LFW was 0.9913 with embedded loss and 0.9727 with a softmax on an L2 comparison function. On the YouTube faces dataset the accuracy and 0.974 with embedded loss and was 0.928 with softmax combined with a L2 comparison function.

The VGG-face algorithm was trained on over 2.6 million face images of celebrities and famous people downloaded from the Web. There were images of 2,622 people with 1,000 image per person. VGG-face was not trained on images in the Notre Dame images collection[2]. The version of the CNN distributed did not use triple embedding during training.

The VGG-face DCNN was modified by removing the last fully connected layer. The output from the penultimate full connected layer was normalized to have a norm of 1. The distance between two images was the L1-norm between their representations produced by the modified DCNN.

IV. Ambient Light Images

Over the last twenty years the most active research area in automatic face recognition has been developing algorithms to recognize faces from frontal still images. In the last ten years, one emphasis of the NIST competitions has been recognition from frontal face images acquired with a digital single lens reflex camera. The majority of these images are considered high quality to humans. The images were collected under two illumination conditions. One was in a studio environment with controlled lighting. The other was under ambient lighting indoors and outdoors.

The Good, Bad, and Ugly (GBU) Face Challenge addressed face recognition with relaxed photometric constraints. Both faces in a pair were acquired in ambient lighting conditions. The images were taken outdoors or indoors in atriums and hallways. To better understand the range of performance under general illumination conditions, three partitions were created based on difficulty of matching two face images of the same person, see Phillips et al. [8] for details. To arrive at the performance-based partitions, three top-performing face recognition algorithms from the FRVT 2006 test were fused to produce a single algorithm. Based on performance of the fusion algorithm, images were divided into three partitions with high (the Good), challenging (the Bad), and very challenging (the Ugly) accuracy, hence the name Good, Bad, and Ugly Face Challenge Problem. In the GBU, the effects of natural variations in a person's day-to-day appearance (hair, facial expression, etc.) and variations in illumination across both indoor and outdoor settings were considered. All of these images were nominally frontal. Because all images were collected between August 2004 and May 2005, aging cannot be a factor. There is the same number of images of each person in all three partitions. Thus, only the images, not the individual identities, changed across the three partitions. This provides an assurance that the accuracy differences were due to factors other than the particular set of face identities tested.

[2]To be precise, we cannot guarantee that no images from Notre Dame dataset were included in the training dataset. This is because the training images were downloaded from the Web. In the worst case, only a few were included, and it should not effect the conclusions in this paper.

Fig. 1. Example of a face-pair of the same person from the GBU partitions. The pair is a very challenging pair.

TABLE I
VR AT FAR = 1 IN 1000 FOR GBU PARTITIONS.

	Ugly	Bad	Good
Fusion baseline	0.15	0.80	0.98
VGG-face	0.26	0.52	0.85

Performance on the GBU was benchmarked by the FRVT 2006 fusion algorithm and the performance metric is the VR at FAR of 0.001. Performance on the GBU partitions for the Fusion baseline and VGG-face are given in Table I and the ROCs[3] are presented in Figure 2. On the Good and Bad partitions, performance on the baseline algorithm is superior. On the Ugly partition, performance of VGG-face is superior. To the author's knowledge, this is the first time an algorithm has achieved better performance than the Fusion baseline.

To gain better understanding of the relative strengths of human performance, Rice et al. [13] examined human performance when algorithms completely fail. From the very-challenging partition in the GBU, 50 same-identity face-pairs and 50 different-identity face-pairs were selected so that the similarity score for all same-identity pairs was lower than all different-identity pairs. A higher similarity scores implies a greater likelihood the face pairs consists of two images of the same face. Thus, performance of the FRVT 2006 fusion algorithm was 100% incorrect, and these are referred to as extremely-difficult face pairs.

To understand the reason for algorithm failure, Rice et al. [13] measured the contribution of face and body, face only, and body only to recognition by humans. To measure the contribution of these three conditions, three versions of the face images were created. In the first experiment,

[3]Because of the very large number of image-pairs in the NIST benchmarks, the error-bars are extremely small and will not show-up on the curves.

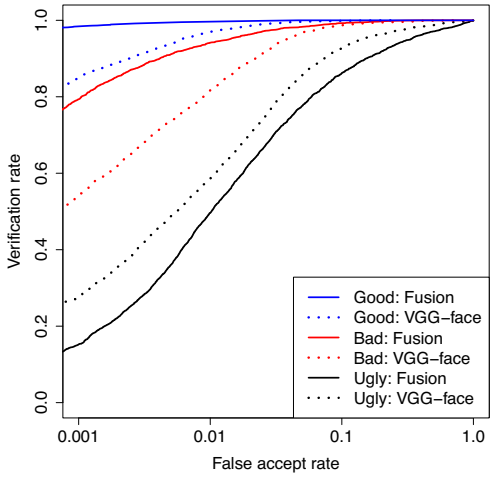

Fig. 2. ROCs on the GBU partitions. ROCs are shown for the baseline Fusion and the VGG-face algorithms.

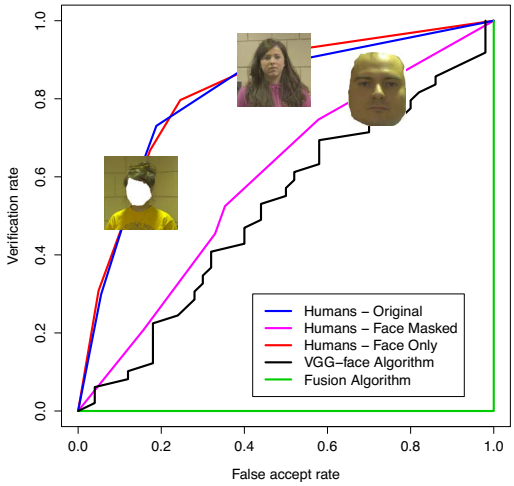

Fig. 3. ROCs on the extremely difficult image-pairs.

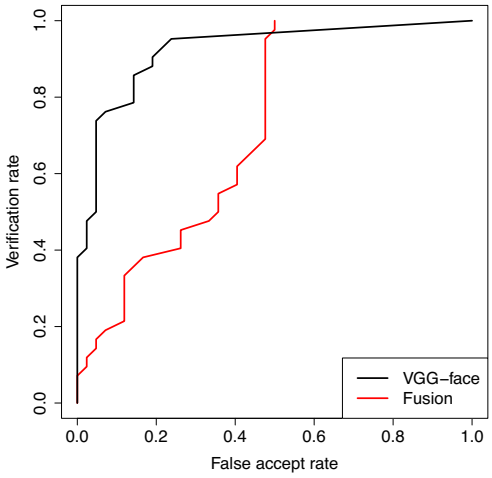

Fig. 4. ROCs for the baseline Fusion and VGG-face on the EFCT. The EFCT was designed to be challenging for facial forensic examiners.

human observers were presented with the original images. In the second experiment, humans were presented with images where the face was masked. In the third experiment, the images consisted of only the face. The ROC for all three human viewing conditions and the VGG-face and the Fusion algorithms are shown in Figure 3.

Performance between the face-masked and original images was not statistically significant, see Rice et al. [13] for statistical analysis. Performance on the face only images was remarkably inaccurate, but greater than chance. The results indicate that the body, rather than the face, accounts for human accuracy at identifying people in the original unedited images. The performance on VGG-face is essentially random and comparable to human performance when only considering the face.

Facial forensic are professionals trained to compare face images. Facial examiners are generally employed by law enforcement and border control agencies. White et al. [14] measured the perceptual accuracy of examiners at face recognition. One standard method for measuring perceptual accuracy is to display two face images side-by-side on a computer screen and ask subjects to rate the similarity of the faces on a five point scale. In White et al. [14], examiners had at most for 30 seconds to view the two faces. In performing their jobs, facial examiners have access to a set of tools. In a perceptual study, the examiners do not have access to their tools and must compare two face images by viewing the images.

To measure the perceptual performance of facial examiners, the Expertise in Facial Comparison Test (EFCT) was created. The EFCT consisted of images from the Bad and Ugly partitions of the GBU. From these two partitions, 84 image pairs were selected based on human accuracy [5][4]. The image pairs were selected to be challenging for untrained people.

Figure 4 shows the ROC for the Fusion baseline and VGG-face on the EFCT. The AUCs for the Fusion and VGG-face algorithms are 0.69 and 0.92, respectively.

On the two easiest conditions, the Good and the Bad partitions analyzed in this section, the Fusion baseline was more accurate than VGG-face; on the the Ugly partition, extremely difficult and EFCT benchmarks, VGG-face was more accurate. Since the release of the GBU face challenge

[4]The EFCT consisted of 168 image pairs with half the image pairs presented upright and half presented in upside down. In the analysis in this paper we only consider the 84 upright pairs.

in 2011, VGG-face is the first algorithm that is more accurate than the Fusion baseline on the Ugly partition.

V. Studio and Ambient Light Images

One focus of the Face Recognition Grand Challenge (FRGC) and the Face Recognition Vendor Test (FRVT) 2006 was recognizing faces when one image was taken in a studio and the other with ambient lighting [10], [12]. The images were full frontal face, and the pictures were taken with a digital single lens reflex camera, see Figure 5.

(a) (b)

Fig. 5. Example of a pair of images used in experiments comparing identities in images captured in a studio environment (a) and an ambient environment (b).

The FRGC is a challenge problem, and the images and supporting data have been available to the face recognition community since 2005. Over the last decade, reported accuracy in the literature has increased to almost perfection. There are two main reasons for the improvement in accuracy. The first is improvements in face recognition algorithms. The second is the optimization of the design of algorithms to the FRGC. To avoid the second reason, the benchmark algorithm is the Viisage submission to the 2006 FRGC workshop. Viisage is a commercial system whose design was not optimized to the FRGC. Thus, the design of both the Viisage system and VGG-face were not optimized for the FRGC. Figure 6 plots the ROC on the FRGC for both VGG-face and the 2006 benchmark[5]. The VRs at a FAR of 0.001 are 0.45 for VGG-face and 0.69 for the benchmark.

The FRVT 2006 was an independent evaluation and algorithms were submitted to NIST for testing. Submissions were tested on sequestered images. Because algorithms were tested on sequestered images, it was not possible to tune the algorithms to the test images. The baseline algorithm for the FRVT 2006 was from Viisage.

Figure 7 plots the ROC on the FRVT 2006 for both VGG-face and Viisage[6]. The VRs at a FAR of 0.001 are 0.52 for VGG-face and 0.79 for the benchmark.

For both the FRGC and FRVT 2006 benchmarks, the baseline algorithms had superior VR at a FAR = 0.001 to VGG-face.

[5]Performance is reported for FRGC version 2, experiment 4, ROC III.

[6]Performance is reported for uncontrolled illumination experiments, Notre Dame dataset, and 1-to-1 protocol.

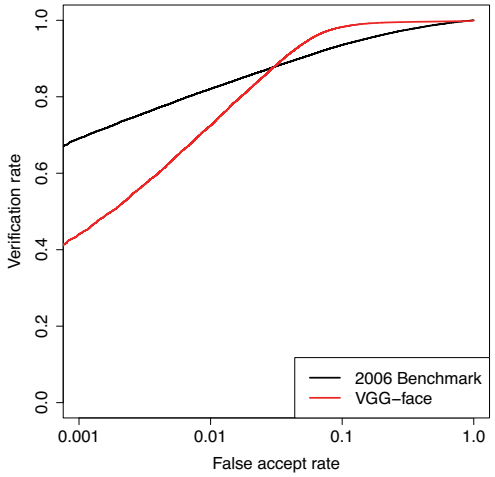

Fig. 6. ROCs for VGG-face and 2006 benchmark on the FRGC.

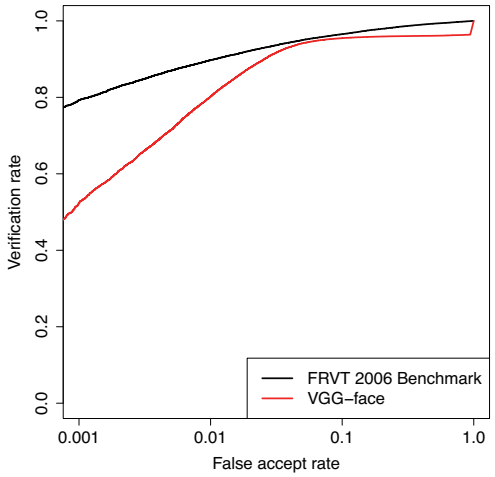

Fig. 7. ROCs for VGG-face and FRVT 2006 benchmark on the FRVT 2006.

VI. Digital Point and Shoot Camera Images

To address and understand the properties of face recognition in unconstrained conditions, the Point and Shoot Face Recognition Challenge (PaSC) was created [1]. The PaSC contains both still images and videos. The images and videos were taken with digital point and shoot cameras, particularly for handheld cameras found in cell phones [1]. By design, the PaSC does not contain images of celebrities. The still image portion consists of 9,376 images of 293 people. Still images were taken at nine locations, both inside buildings and outdoors, with five point-and-shoot still cameras. Still images were taken at a variety of poses and distances from the camera see Figures 8 and 9.

Fig. 8. Example of one block of still PaSC images. The block contains four frontal images, four non-frontal images, four images close to the camera, and four images far from the camera. Courtesy of Beveridge et al. [1].

Fig. 9. Cropped face images extracted from still images in the PaSC. These images demonstrate some of the complications that arise in point-and-shoot images, lighting, motion blur and poor focus. Courtesy of Beveridge et al. [1].

Performance analysis was limited to the still frontal images taken with the hand-held cameras. Performance on the PaSC was benchmarked by a PittPatt SDK 5.2.2-based face recognition algorithm. On the still images, accuracy of the PittPatt algorithm is the best in the literature. Figure 10 plots the ROC on the still frontal face images in the PaSC for both VGG-face and the PittPatt algorithm. The VRs at a FAR of 0.001 are 0.50 for VGG-face and 0.34 for the benchmark. On the PaSC benchmark, VGG-face had superior VR at a FAR of 0.001.

VII. Human Performance

Since the FRGC, human performance has been systematically included in the NIST face recognition benchmarks [6]. Measuring human performance allows for a direct comparison of human and algorithm accuracy. Phillips and O'Toole [11] present an overview of these studies and this analysis follows the methods in this paper. We compare human and VGG-face accuracy on the FRGC, the FRVT 2006, the GBU, extremely difficult faces, and EFCT [6], [12], [5], [13], [14].

Performance was measured by presenting two face images on a computer screen. Subjects were asked to judge the similarity between two faces on the following 5 point scale:

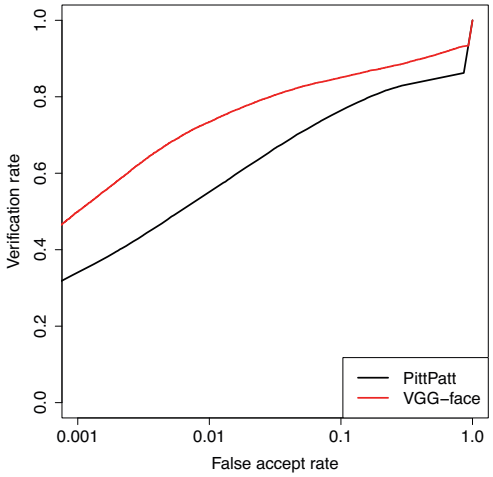

Fig. 10. ROCs for VGG-face and PittPatt benchmark on the still frontal images in PaSC.

1. Sure they are the same,
2. Think they are the same,
3. Don't know,
4. Think they are not the same, and
5. Sure they are the same.

From the human generated ratings, ROCs were computed. For each experiment, ratings were collected on at most 240 pairs of faces with half the pairs having the same identity and the other half having different identities. The selection criteria for the face-pairs varied by experiment and details can be found in the references. All the experiments were conducted in a laboratory setting and none of the experiments depended on crowd sourcing methods.

With the exception of the EFCT, human performance was measured on normal (i.e., untrained) people who had no professional experience with face recognition. The EFCT measured the perceptual performance of facial examiners. For each experiment, accuracy is summarized by the AUC. Figure 11 compares human and VGG-face accuracy across seven experiments. There are two experiments on the FRGC: easy and difficult. The AUC for VGG-face on the extreme difficult benchmark is 0.5.

With exception of the extremely difficult and the EFCT benchmarks, the VGG-face algorithm's performance was superior to humans. The EFCT reports accuracy for facial forensic examiners and the accuracy is comparable to VGG-face. On the EFCT, the AUC for normal people is 0.84.

For the face-pairs in the extremely difficult benchmark, there are minimal identity cues in the interior of the face. The identity cues are in the body, see Figure 3. Human performance on the interior of face is essentially random and comparable to VGG-face. The results on the seven experiments in this analysis suggest that VGG-face is extracting more identity information out of the face than untrained humans.

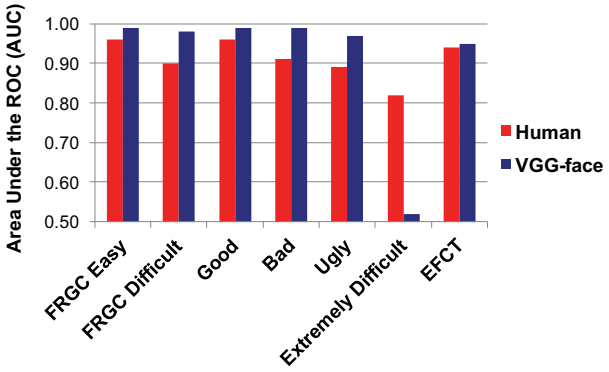

Fig. 11. Summary of human and algorithm performance across benchmarks.

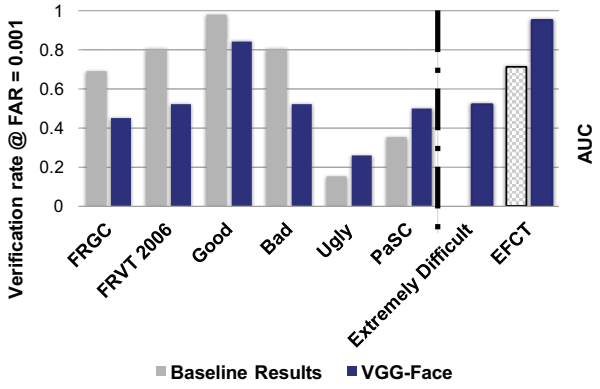

Fig. 12. Summary of VGG-face performance and existing performance on eight NIST face-recognition benchmarks. For the benchmarks to left of the dashed, VR at a FAR of 1 in 1000 is reported; for the benchmarks to right, AUC is reported. On the extremely difficult benchmark, the existing algorithm AUC is 0.0. The existing results bars are the performance for the baseline algorithms discussed in the paper.

The results on the EFCT suggest that further experiments should investigate if VGG-face is comparable to experts.

VIII. Summary of Performance and Conclusions

The VGG-face algorithm was designed to recognize faces in variable illumination. This is one of the properties that characterize the four hardest benchmarks in our study. Figure 12 summarizes the accuracy VGG-face performance on the NIST benchmarks. In addition, for each benchmark, we report the accuracy of the baseline algorithms discussed in this paper (Existing Results in Figure 12). For the six benchmarks to the left of the dashed line, we report the VR at a FAR of 1 in 1000; for the two benchmarks to the right, we report AUC. On the FRGC, the FRVT 2006, the Good and the Bad benchmarks, algorithms from 2006 were more accurate. One common characteristic across these four benchmark is reduced variability in illumination among the images.

To gauge the performance of algorithms over a variety of conditions, we recommend that performance of face recognition algorithms should be characterized by a set of benchmarks. This would encourage the development of algorithms that are robust to different scenarios. It may not be possible to achieve optimal performance for all scenarios. In this case, performance on multiple benchmarks would show the trade-offs in optimizing accuracy for a single benchmark. A set of benchmarks would lead to a better understanding of human and algorithm performance. The knowledge of the relative performance of humans and algorithms could lead to more efficient division of duties between algorithms and humans in face recognition systems.

References

[1] J. Beveridge, P. Phillips, D. Bolme, B. Draper, G. Givens, Y. M. Lui, M. Teli, H. Zhang, W. Scruggs, K. Bowyer, P. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8, Sept 2013.

[2] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[3] 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.

[4] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, 2014.

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

[6] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Pénard, and H. Abdi. Face recognition algorithms surpass humans matching faces across changes in illumination. *IEEE Trans. PAMI*, 29 1642-1646:1642–1646, 2007.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *Proceedings of the British Machine Vision*, 2015.

[8] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. An introduction to the Good, the Bad, and the Ugly Face Recognition Challenge problem. In *Proceedings Ninth IEEE International Conference on Automatic Face and Gesture Recognition*, 2011.

[9] P. J. Phillips, P. J. Flynn, and K. W. Bowyer. Lessons from collecting a million biometric samples. *Image and Vision Computing*, 2016.

[10] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the Face Recognition Grand Challenge. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 947–954, 2005.

[11] P. J. Phillips and A. J. O'Toole. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(74-85), 2014.

[12] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results. *IEEE Trans. PAMI*, 32(5):831–846, 2010.

[13] A. Rice, P. J. Phillips, V. Natu, X. An, and A. J. O'Toole. Unaware person recognition from the body when face identification fails. *Psychological Science*, 24:2235–2243, 2013.

[14] D. White, P. J. Phillips, C. A. Hahn, M. Q. Hill, and A. J. O'Toole. Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B*, 282, 2015.

[15] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534, 2011.