A Comparison of Human and Automated Face Verification Accuracy on Unconstrained Image Sets*

Austin Blanton
Noblis
imaus10@gmail.com

Kristen C. Allen
Noblis
kristen.allen@noblis.org

Tim Miller
Noblis
timothy.miller@noblis.org

Nathan D. Kalka
Noblis
nathan.kalka@noblis.org

Anil K. Jain
Michigan State University
jain@cse.msu.edu

Abstract

Automatic face recognition technologies have seen significant improvements in performance due to a combination of advances in deep learning and availability of larger datasets for training deep networks. Since recognizing faces is a task that humans are believed to be very good at, it is only natural to compare the relative performance of automated face recognition and humans when processing fully unconstrained facial imagery. In this work, we expand on previous studies of the recognition accuracy of humans and automated systems by performing several novel analyses utilizing unconstrained face imagery. We examine the impact on performance when human recognizers are presented with varying amounts of imagery per subject, immutable attributes such as gender, and circumstantial attributes such as occlusion, illumination, and pose. Results indicate that humans greatly outperform state of the art automated face recognition algorithms on the challenging IJB-A dataset.

1. Introduction

Human capabilities have served as the "gold standard" for facial recognition [19]. Barring knowledge of the circumstances of capture, human comparison of facial imagery is the best way to verify or differentiate faces within imagery or video, and has been successfully used in the creation of testing databases at scale [8]. As automated recognition approaches human levels of accuracy, the knowledge that humans can recognize faces across variations in age, pose, illumination and expression (A-PIE) requires additional improvement at such comparisons. The common goal for computer vision systems is to match or surpass human accuracy at the same task with objectivity and efficiency. In the context of automated face recognition on constrained images, this gap has been bridged. Recent publications on faces in the wild benchmarks such as LFW [7] and YTF [21] have seen this gap narrow as well [3].

In this work, we seek to expand on previous comparisons [2, 5, 12, 17, 13, 16] of the recognition accuracy of humans and automated systems by performing three novel analyses: (i) measuring and comparing accuracy on unconstrained face imagery, (ii) measuring accuracy when comparing image sets (as opposed to single image comparisons) and (iii) examining the effect of immutable attributes such as gender and circumstantial attributes such as illumination, occlusion and pose. The research is enabled by the recently released IJB-A dataset [8], which contains media in the wild images and videos from 500 subjects, with full pose variation. This is notably different from the seminal LFW dataset, and other media in the wild datasets, which were filtered to only contain faces detectable by a commodity face detector trained on frontal faces.

Through these analyses, we seek to understand whether or not human accuracy degrades in a similar manner as machines when the quality of a face sample degrades (e.g., extreme pose, adverse illumination, occlusion). It is known that algorithms perform worse on fully unconstrained face images [8, 20] but, to our knowledge, such a study has not yet been performed for human recognition. Because IJB-A is the first unconstrained face dataset unbiased by detectable faces, it is important to establish a human baseline for automated methods to aspire to match and surpass humans.

*This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via FBI Contract # GS10F0189T-DJF151200G0005824. 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.

Figure 1: Example faces from the FERET [15], LFW [7], and IJB-A [8] datasets.

Similarly, we seek to understand whether multiple samples (instances) from a subject can compensate for the effects of large pose variations.

2. Related Work

There have been several studies [2, 5, 12, 17, 13, 16] analyzing the performance of humans and algorithms on the face verification task: two face images are shown side by side (referred to as a pair), and the human or algorithm decides if they are the same identity or not. These studies have predominantly used frontal face imagery, i.e. FERET [15] and FRVT [6].

Recent studies that compare human performance to algorithm performance on the face recognition task note the other-race effect. The other-race effect is observed when individuals of a particular race perform better at identifying individuals of their own race than a different race. This effect has been observed in humans since the 1970s [5], and in a number of algorithms [5, 12, 17, 13, 16]. It is believed that the dataset used to train the algorithm is part of the reason for this effect. Results from these studies illustrated that when the algorithm originates in a Western country, the algorithm performs better on Western faces than East Asian faces, but when the algorithm was developed in an East Asian country (presumably trained predominantly on Asian faces), it performs better on East Asian faces. In an auto-associative neural network [12], the exposure of statistical learning algorithms to other-races impacted the feature space for representation. In [9], Klare et al. evaluated six face recognition algorithms on a database subdivided into three demographic categories including race (Black, Hispanic and White), sex, and age (younger, middle-aged and older). All recognition algorithms performed worse on the Blacks, females and younger subject demographic groups. The role of training was also studied by comparing the performance of the 4SF algorithm when it was trained from all three ethnic groups simultaneously to implementations of the same algorithm trained with each of the ethnic groups individually. The highest recognition accuracy was achieved when the system was trained only on faces of the same ethnicity.

In [2], human performance on unconstrained still-to-still and video-to-video face matching scenarios is explored. Human's perform better than face recognition algorithms when performing the task of matching unconstrained faces. Furthermore, Best-Rowden et al. [3] illustrated that fusion of human confidence and automated FR match scores improves the overall recognition performance, suggesting that both offer complementary information for matching unconstrained faces. A fusion of algorithm performance and human performance using the partial least squares regression method, was found to improve performance of the human ID task over algorithm or human performance by itself [10]. Other studies on the human recognition task include both other-race faces and engineered illumination differences [1, 4, 11, 17, 14, 11]. Image pairs used were rated as poor, moderate, and good, based on the strength of their similarity scores (a high similarity score assigned the corresponding pair a good rating, and a low similarity score assigned a poor rating). The results of these studies depend on the level of difficulty of the face pair; as the difficulty of the image pair increases, the algorithm performance becomes more similar to that of human performance [11]. Specifically, O'Toole et al. found that out of seven algorithms, three performed better than humans on what would be considered poor face pairs, while all but one performed better on good face pairs [14]. It is important to note that these results were established using faces that were unfamiliar (not known) to the humans.

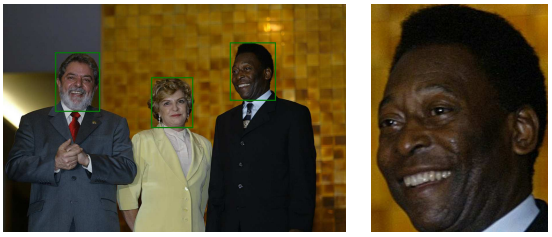
Prior work has predominately focused on studying the other race effect and comparing human to automated algorithm performance utilizing frontal face imagery. In contrast, there have been relatively few studies on comparisons between human and automated facial recognition performance when utilizing facial imagery from fully unconstrained benchmarks, which is the primary focus of this paper.

3. Experimental Design

The following experimental methodologies were designed to compare human performance to state-of-the-art automated facial recognition methods.

3.1. Protocol

The IJB-A dataset was collected from creative commons images on the Internet. The face and landmark locations were manually annotated by MTurk workers (Figure 2). IJB-A has 500 subjects, each with an associated geographic origin, skin tone on a scale from one to five, and gender. A breakdown of the subject demographics is shown in Figure 3. Additionally, IJB-A provides crowd-sourced values for age, facial hair, lighting, and occlusion categories, as well as a pose estimate from PittPatt 5 (PP5), where available.



(a) (b)

Figure 2: An example of face bounding boxes from the IJB-A dataset. (a) The bounding boxes overlaid on the image. (b) Example face image cropped to the bounding box.

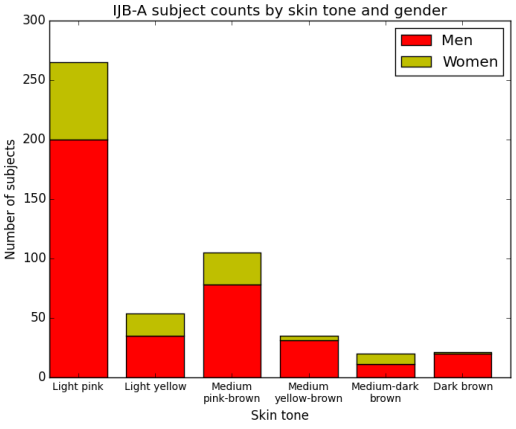


Figure 3: Demographic breakdown of subjects in the IJB-A dataset. There is a much greater percentage (75%) of males than females.

For each subject with a sufficient no. of images to create multiple probe templates, there are two genuine self-comparisons. Of the two, one comparison has a single image in either the probe or gallery and the other comparison has multiple images in both probe and gallery. For the multiple images comparisons, the number of images in probe and gallery is randomly selected per-comparison. In addition to the genuine comparisons, there are approximately $10,000$ impostor comparisons. Impostor comparisons are within two skin tone gradations from each other and share the same gender. The impostor comparisons are composed of $5,000$ random single image comparisons and $5,000$ multi-image comparisons. See Table 1 for a compact representation of the protocol for comparisons.

	# of comparisons	Comparison types
Genuine	976 *(2 per subject)*	488 (single-image) 488 (multi-image)
Impostor	10,000	5,000 (single-image) 5,000 (multi-image)

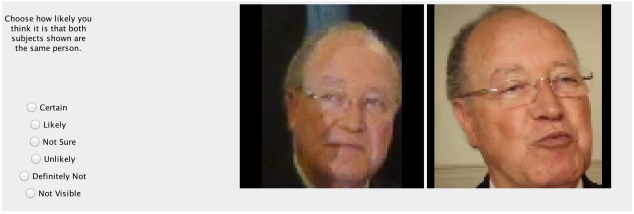
Table 1: Breakdown of the protocol for comparisons. Single-image comparison means at least one side of the comparison (probe or gallery) has only a single image. Multi-image means both sides (probe and gallery) of the comparison have at least two images.

3.2. Crowd Sourced Verification

Amazon Mechanical Turk (MTurk) workers are presented with face images of two subjects that may or not be the same (Figure 4). The number of images of each subject in the comparison can be anywhere from one to six. The workers select how sure they are the two subjects have the same identity on a scale of one to five, with one being certain they are the same, three being unsure, and five being certain they are not the same. A sixth option, "Not Visible", is used to indicate if a face is not visible in the image. The scale is presented to the worker with the following wording: {Certain, Likely, Not Sure, Unlikely, Definitely Not, Not Visible}.

Ten MTurk workers are assigned to each comparison, five in the US and five in India. If the majority of workers agree that either subject is not visible, the comparison is not used for analysis. In practice, it was found that often workers chose the "Not Visible" option for images where a face is clearly visible, but at low resolution (see Figure 5b), as opposed to images where no face is visible (Figure 5a). Thus, any remaining not visible answers are changed to "Not Sure". All worker answers are averaged to produce a human match score.

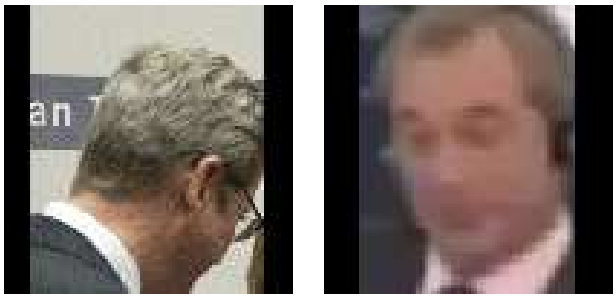
Because the actual identity of each subject in a comparison is known, the quality of each worker can be evaluated relative to the rest of the workers. Therefore, if a particular worker makes too many mistakes, their work can be



(a)

(b)

Figure 4: Two examples of the interface presented to MTurk workers. (a) A genuine single-image comparison. (b) An impostor multi-image (6 images v. 4 images) comparison.



(a) (b)

Figure 5: Two examples of images marked "Not Visible" by MTurk workers.

rejected. Figure 6 shows the distribution of worker accuracy, where a correct answer is either a 4 or 5 response in the case of genuine comparisons and a 1 or 2 response in the case of impostor comparisons. An answer of 3 ("not sure") is not counted and all other cases are considered incorrect. A worker's accuracy is the percentage of correct answers. Most workers are quite accurate, as the histogram in Fig. 6 illustrates a skew toward accuracies above 90%. This indicates that even for the challenging images in IJB-A, face verification is an easy task for humans. If we view Figure 6 as the left tail of a Gaussian curve with $\mu = 1.0$, selecting a rejection threshold on the left tail that is more than two standard deviations from the mean ensures that only workers whose answers are a statistical outlier relative to the population are rejected. As 60% accuracy is clearly at the tail end of the curve, all answers from workers having accuracy less than or equal to 60% are removed. A score of 60% is 2.47 standard deviations from μ. Answers from users with accuracy greater than 60% account for 96.8% of all data. Ultimately, there were 414 unique workers of quality used.

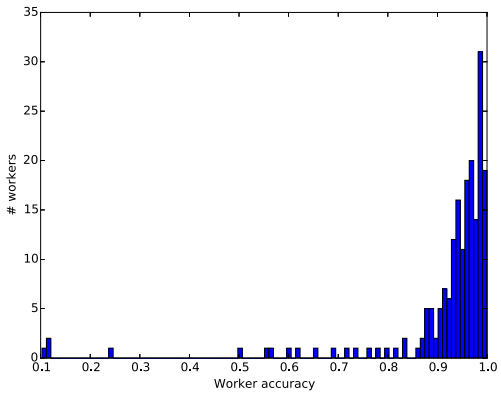


Figure 6: Histogram of worker accuracies. All answers from workers with accuracy lower than 60% were rejected.

3.3. Experiments

In order to compare human face recognition to automated methods, each comparison was input to two face recognition algorithms: PittPatt5 (PP5) and a state of the art convolutional neural network (FR-CNN) implementation [18]. The CNN architecture was trained utilizing imagery from the publicly available CASIA webfaces database [22]. Our implementation has an accuracy of 93.56% (at 0.1% FAR) for the standard LFW protocol which is lower than best reported accuracy of 99.77%. We attribute this difference in performance to using a smaller training set and not fine-tuning with triplet loss.

- **Comparison to human accuracy**. To produce an overall match score for the two multi-instance templates, each of the first subject's images are compared to each of the second subject's images. All the scores are averaged and can then be compared to the average MTurk worker score.
- **Immutable attributes**. The other-race effect, noted in both human and algorithm performance, is posited to be caused in both cases by a bias in the types of faces experienced [12]. To study if a similar bias exists for the IJB-A dataset, we study male v. male and female v. female, and compare the results among these two categories.

- **Multi-instance comparisons**. To investigate if multiple images per subject improves recognition, the comparisons were separated into categories of single-image comparisons, two-image comparisons, etc., up to six-image comparisons. The expectation is that human performance increases as the minimum number of images per subject increases.

- **Circumstantial attributes**. In the same vein, we examine the effect of circumstantial attributes (lighting, occlusions and pose) on verification. Because the comparisons have multiple images, often with a possibility of errors in these attributes, only single-image comparisons are used. In this way we isolate the effect of variations in the attribute - human recognition is limited to a single attribute value for one of the subjects.

4. Results

Figure 8 shows the distribution of worker match scores, color-coded as genuine or impostor comparisons. There is a clear separation between genuine and impostor scores, with very few "Not Sure" answers and few false positive and false negative outliers. The majority of answers indicate absolute certainty of the subject identities. Compare this to the distribution for the automated algorithms. Figure 7 shows that humans are quite good at recognizing faces in challenging images such as those in IJB-A; they outperform algorithms (in terms of TAR) at every false accept rate by a large margin.

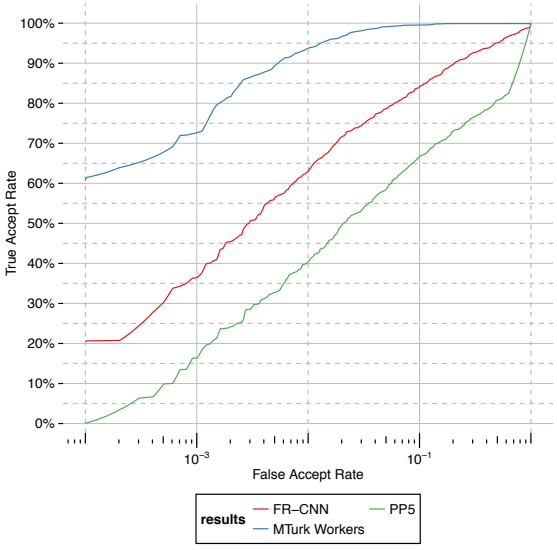


Figure 7: Humans outperform the two face recognition algorithms by a large margin on challenging images such as those in IJB-A.

In studying the human performance for immutable attributes, it was found that MTurk workers are able to recognize male faces more easily than female faces (figure 9). The gap, while not very pronounced in human performance, is wider for the automated methods. However, this may be more indicative of a greater percentage of male subjects in IJB-A - 75% of the subjects are male. Additionally, there may be an unknown bias in the workers that predisposes them toward recognizing male faces.

Figure 10 shows that the number of images per subject in the comparison does have an effect on human face recognition performance. However, the relationship does not appear to be exactly linear. Comparisons with a subject having only a single image are clearly harder for humans and those having at least six images for both subjects are easier, but the number of images in between 1 and 6 are relatively similar. In other words, if the number of images per subject exceeded a threshold, it leads to "good enough" performance.

Other circumstantial artifacts of the image, such as lighting, occlusions, and pose, are hard problems for automated methods. Figure 11 shows that while humans perform similarly for both indoor and outdoor images, algorithms tend to work better on indoor images. Figure 12 confirms that the issues facing state of the art algorithms with pose are also problems for human recognition. As pose gets increasingly off-center, recognition performance decreases. Figure 13 reveals that perhaps the nose and mouth are very important regions for human face recognition - the performance for images where the mouth or nose are occluded is noticeably worse.

5. Conclusion

In this work, we compared the recognition accuracy of humans and automated state of the art facial recognition systems. A key distinction between this study and previous studies involving such comparisons is that this study utilized fully unconstrained face imagery as reflected in the IJB-A dataset. We performed several novel analyses: (i) measured and compared accuracy on fully unconstrained face imagery, (ii) measured accuracy when comparing multiple image sets (as opposed to single image comparisons), and (iii) examined the effect of immutable attributes such as gender and circumstantial attributes (lighting, occlusions, pose) on human verification. Overall, our results showed that humans overwhelmingly outperformed automated algorithms on the challenging images in the IJB-A dataset.

For immutable attributes such as gender, our results suggest that males are easier to recognize than females for humans. This trend was observed to be more pronounced for automated algorithms. However, it is important to note that caution must be used interpreting this observation as it is confounded by a bias in the fraction of males (75%) and females within IJB-A.

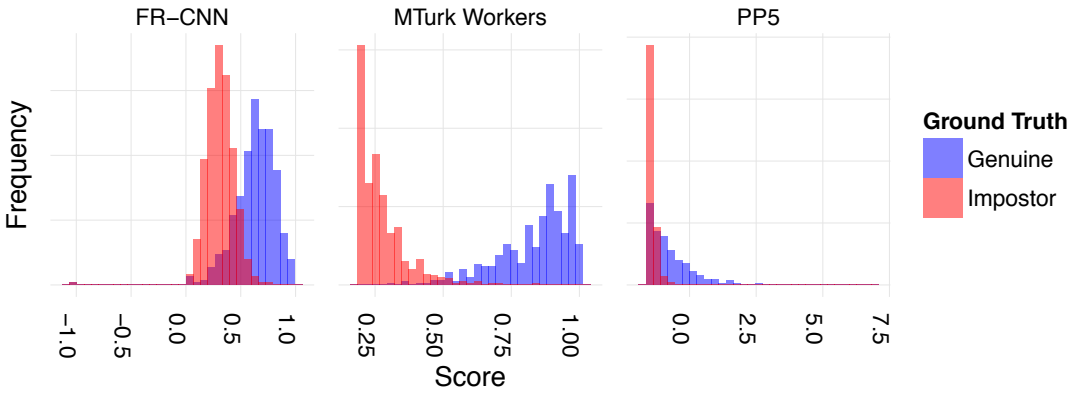


Figure 8: The distribution of genuine/impostor scores is mostly separable for human verification, which indicates the human ability to correctly recognize faces. The two automated algorithms have greater overlap in the genuine and imposter distributions and perform correspondingly worse.

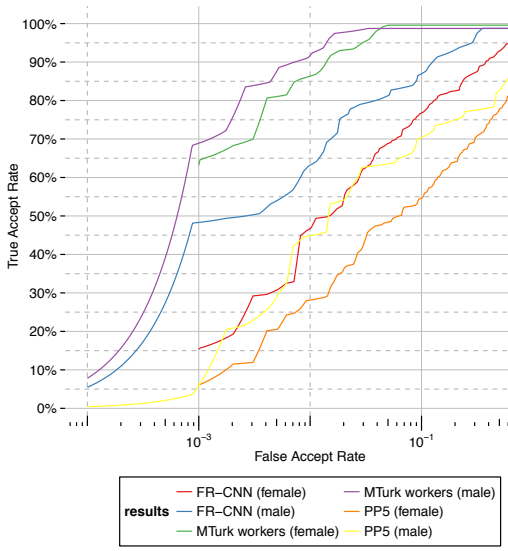


Figure 9: Human and automated recognition separated by gender. Notice that the gap between male and female accuracy is smaller for MTurk workers than for automated methods.

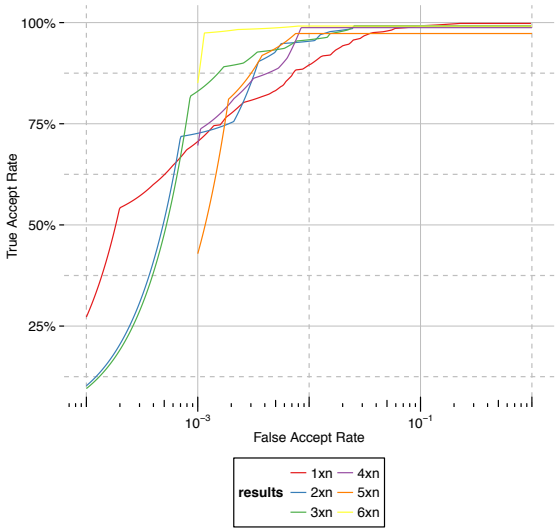


Figure 10: The minimum number of images per subject in the comparison does have an effect on human performance. In particular, comparisons with a subject having only one image are clearly harder for humans and those having at least six images for both subjects are easier. Here n represents the no. of images on the right half of a multi-image comparison (illustrated in Fig. 4b).

With regard to face image set comparisons, subjects having only a single image are significantly more difficult for humans to recognize than those having at least six images for both subjects. Our experiments studying circumstantial subject attributes illustrate that (i) humans perform similarly for both indoor and outdoor environments while outdoor environments are more challenging for automated algorithms, (ii) facial occlusion of the nose and mouth degrade human recognition performance, and (iii) extreme pose presents a problem for both human and automated recognition algorithms.

Future work will further study the influence of the other race effect on the training of automated face recognition algorithms. Beyond that, those instances within the uncon-

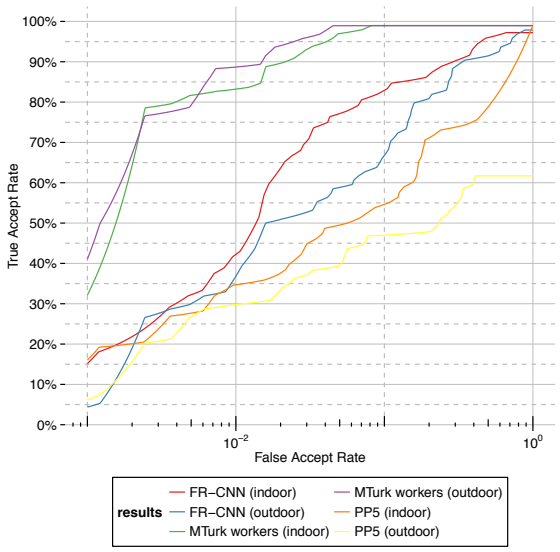


Figure 11: Human and automated recognition on indoor and outdoor images. While humans show little difference between the two, the algorithms studied have a noticeable gap, performing better on indoor images.

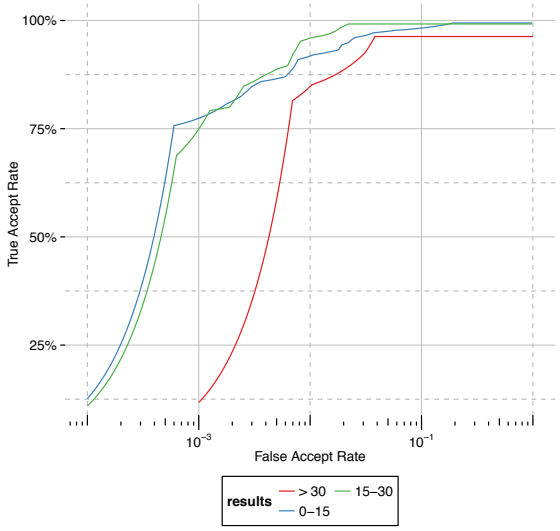


Figure 12: Human recognition on various pose categories, as estimated by PP5 (in degrees). Humans have more difficulty identifying off-pose faces, with the degree of difficulty increasing with the increase in yaw angle.

strained face datasets that are exceptionally difficult for human recognition will be explored which is likely to guide further developments in automated algorithms towards the goal of surpassing human performance.

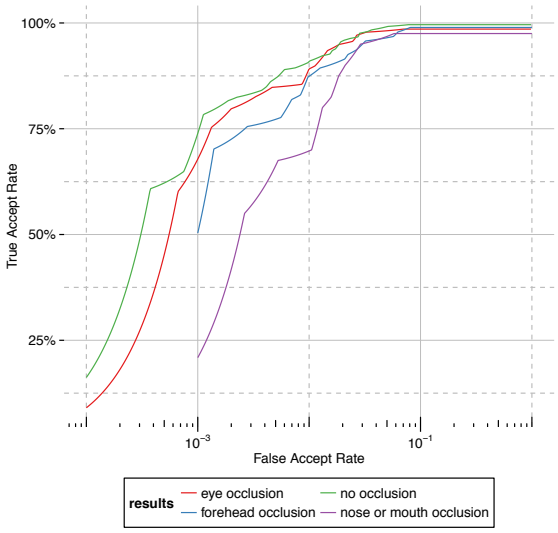


Figure 13: Human recognition on images with various facial features occluded. Images with the nose or mouth occluded are the most difficult for humans to identify. This indicates that humans likely rely heavily on information in that region of the face.

References

[1] A. Adler and M. Schuckers. Comparing human and automatic face recognition performance. *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1248–1255, Oct 2007. 2

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

[3] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Trans. on Information Forensics and Security*, 9(12):2144–2157, Dec 2014. 1, 2

[4] V. Bruce, P. Hancock, and A. Burton. Comparisons between human and computer recognition of faces. In *Proc. Third IEEE International Conf. on Automatic Face and Gesture Recognition*, pages 408–413, Apr 1998. 2

[5] N. Furl, P. J. Phillips, and A. J. OĂŹToole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26(6):797 – 815, 2002. 1, 2

[6] P. Grother and M. Ngan. Face recognition vendor test (frvt): Performance of face identification algorithms. Technical Report 8009, National Institute of Standards and Technology, 2014. 2

[7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Re-

port 07-49, University of Massachusetts, Amherst, October 2007. 1, 2

[8] B. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1931–1939, June 2015. 1, 2

[9] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, Dec 2012. 2

[10] A. O'Toole, H. Abdi, F. Jiang, and P. Phillips. Fusing face-verification algorithms and humans. *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1149–1155, Oct 2007. 2

[11] 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. Appl. Percept.*, 9(4):16:1–16:13, Oct. 2012. 2

[12] A. J. O'Toole and V. Natu. Computational perspectives on the other-race effect. *Visual Cognition*, 21(9-10):1121–1137, 2013. 1, 2, 4

[13] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop. Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, 30(3):169 – 176, 2012. 1, 2

[14] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi. Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(9):1642 – 1646, Sept. 2007. 2

[15] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, Oct 2000. 2

[16] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole. An other-race effect for face recognition algorithms. *ACM Trans. Appl. Percept.*, 8(2):14:1 – 14:11, January 2011. 1, 2

[17] P. J. Phillips and A. J. O'Toole. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(1):74 – 85, 2014. 1, 2

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. 4

[19] E. Taborsky, K. Allen, A. Blanton, A. Jain, and B. Klare. Annotating unconstrained face imagery: A scalable approach. In *International Conf. on Biometrics*, pages 264–271, May 2015. 1

[20] D. Wang, C. Otto, and A. K. Jain. Face search at scale: 80 million gallery. Technical report, Michigan State University, 2015. 1

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

[22] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. 4