To appear in *Proc. IEEE IJCB, 2014*

Unconstrained Face Recognition: Establishing Baseline Human Performance via Crowdsourcing

Lacey Best-Rowden,[1] Shiwani Bisht,[2] Joshua C. Klontz,[3] Anil K. Jain[1]
[1]Michigan State University, East Lansing, MI, U.S.A.
[2]Cornell University, Ithaca, NY, U.S.A.
[3]Noblis, Falls Church, VA, U.S.A.
bestrow1@cse.msu.edu;sb854@cornell.edu;joshua.klontz@noblis.org;jain@cse.msu.edu

Abstract

Research focus in face recognition has shifted towards recognition of faces "in the wild" for both still images and videos which are captured in unconstrained imaging environments and without user cooperation. Due to confounding factors of pose, illumination, and expression, as well as occlusion and low resolution, current face recognition systems deployed in forensic and security applications operate in a semi-automatic manner; an operator typically reviews the top results from the face recognition system to manually determine the final match. For this reason, it is important to analyze the accuracies achieved by both the matching algorithms (machines) and humans on unconstrained face recognition tasks. In this paper, we report human accuracy on unconstrained faces in still images and videos via crowdsourcing on Amazon Mechanical Turk. In particular, we report the first human performance on the YouTube Faces database and show that humans are superior to machines, especially when videos contain contextual cues in addition to the face image. We investigate the accuracy of humans from two different countries (United States and India) and find that humans from the United States are more accurate, possibly due to their familiarity with the faces of the public figures in the YouTube Faces database. A fusion of recognitions made by humans and a commercial-off-the-shelf face matcher improves performance over humans alone.

1. Introduction

Automatic face recognition systems are currently deployed in many applications, including mobile device authentication, identity card de-duplication, and security portal verification. Face recognition technology also has the potential to aid law enforcement in situations such as watch-list surveillance or forensic identification scenarios like the Boston Marathon bombings [8]. However, this capability has not yet been realized due to the intrinsic difficulties in matching low-quality face images and videos present in unconstrained environments (Fig. 1). Consequently, face recognition in law enforcement scenarios generally involves a "human in the loop," where frames of interest are extracted and the top matching candidate faces are manually adjudicated [7].

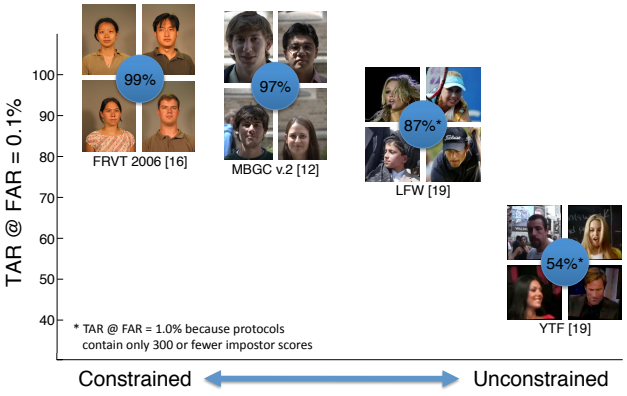


Figure 1. Accuracies of face recognition systems degrade in unconstrained matching scenarios. FRVT 2006 consists of frontal face images with controlled illumination and neutral expressions [12]; MBGC v.2 Controlled vs. Uncontrolled challenge problem contains frontal face images with variations in illumination and expressions [12]; LFW [6] and YTF [21] databases consist of face images and videos, respectively, with arbitrary pose, illumination, expression, and occlusion.[1]

Large legacy face databases, *e.g.* driver's licenses and mug shot photos, have historically been the focus of face recognition researchers as a still-image matching problem. Given the extremely high recognition accuracies in NIST evaluations on such face images [16], there is now a growing interest in face matching algorithms that operate on more challenging scenarios, including face recognition in

[1]Reported performances are the best published results as of July 2014.

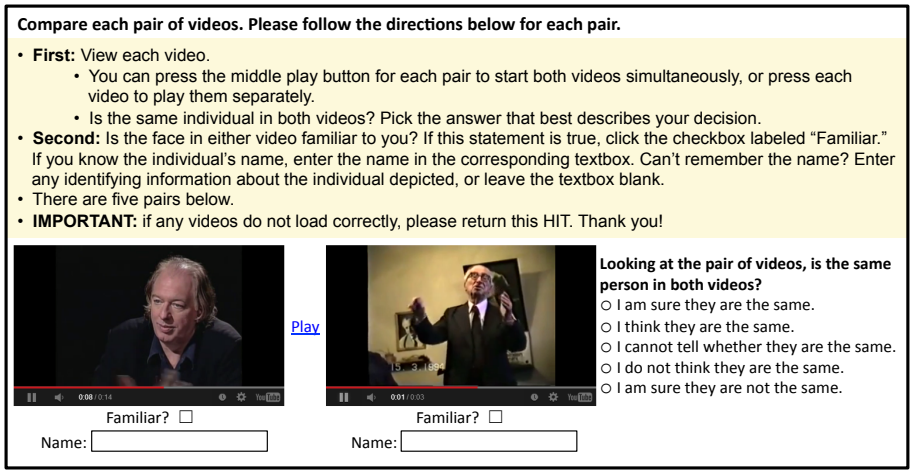


Figure 2. The MTurk interface used to measure human face recognition performance. The example demonstrates the video-to-video face verification task on the YTF database. An analogous interface was used for still-to-still face verification on the LFW database.

videos. This may be attributed to the ubiquity of CCTV and mobile device video footage in conjunction with emerging forensic and security scenarios in urgent need of automated face recognition. For example, a robber, caught on surveillance cameras in 2013, was recently identified by Chicago police using face recognition technology.[2] The 2011 Vancouver Stanley Cup riots and 2013 Boston Marathon bombings are also examples of criminal or terrorist incidents where a mature face recognition technology for unconstrained face matching in videos could have been valuable to law enforcement agencies.

There exist several public domain databases designed to study unconstrained face recognition. Two noteworthy collections are the Labeled Faces in the Wild (LFW) [6] and YouTube Faces (YTF) [21] databases, published to challenge computer vision researchers with large-scale unconstrained face recognition in still images and videos, respectively.[3,4] Understanding how humans are able to accurately recognize faces in unconstrained environments (*e.g.* the LFW and YTF databases) may offer insight into why algorithms are not yet able to achieve the same robustness in the presence of confounding visual factors.

In this paper, we present human performance, measured via crowdsourcing on Amazon Mechanical Turk (MTurk),[5] for face verification tasks on still-to-still and video-to-video face matching on the LFW and YTF databases, respectively (Fig. 2). The key contributions of this paper are: (i) Confirm the human accuracy results reported in [10] on the LFW database and show that humans can verify the most difficult LFW face pairs when presented with two images per subject; (ii) *establish the first human baseline on the YTF video database*; (iii) show that human performance on the YTF database depends on the nationality of the workers in the crowdsourcing study and their familiarity with the faces in question; (iv) compare face recognition performance by humans and state-of-the-art algorithms; (v) provide insight into the discrepancy between false matches made by humans and algorithms on the LFW and YTF databases; and (vi) demonstrate that fusion of human and algorithm recognition results can improve performance for both still-to-still and video-to-video face matching.

2. Related Work

O'Toole *et al*. studied human versus algorithm performance for face verification on a portion of the Face Recognition Grand Challenge (FRGC) image database [14]. Average human performance from 49 individuals was compared against seven different algorithms on 240 face image pairs empirically determined as either "easy" or "difficult" with regards to illumination conditions. The authors found that six of the seven algorithms outperformed humans on the easy pairs, but only three algorithms were more accurate than humans on the difficult pairs. However, the number of face image pairs in this study was relatively small, and the faces exhibited little to no variation in pose, ethnicity, or aging between photographs. In fact, in a separate study by O'Toole *et al*. on the Good, the Bad, and the Ugly challenge problem [13], face recognition algorithms outperformed humans on frontal face verification tasks but were less adept at overcoming pose variations.

Kumar *et al*. conducted a study using MTurk to measure human performance on the LFW database. The authors reported three different face verification tasks on (i) the orig-

[2] http://www.suntimes.com/27895985-761/armed-robber-identified-by-facial-recognition-technology-gets-22-years.html
[3] http://vis-www.cs.umass.edu/lfw/
[4] www.cs.tau.ac.il/~wolf/ytfaces/
[5] www.mturk.com/

Study	Database	Scenario	Key Findings
Adler and Schuckers [1]	NIST Mugshot*	Still-to-still	Only 30% of experiment participants outperformed the best algorithm tested.
O'Toole *et al*. [14]	FRGC*	Still-to-still	Six (three) of the seven algorithms in the study surpassed humans for "easy" ("difficult") face image pairs. Difficulty was determined by the matching scores of a PCA algorithm. All face images were frontal with variations in illumination.
Kumar *et al*. [10]	LFW	Unconstrained still-to-still	Human accuracy on face image pairs was 99.20%. Recently published algorithm accuracies are comparable: 97.35% [19] and 97.45% [18].
Phillips *et al*. [16]	FRVT 2006*	Still-to-still	Out of the seven algorithms tested, one outperformed humans and two were comparable.
Chen *et al*. [5]	FOCS UTD	Unconstrained still-to-still and video-to-video	Algorithms were comparable to humans on face only videos for scenarios with limited pose variations, but humans outperformed algorithms on faces and bodies and cross pose scenarios.
O'Toole *et al*. [13]	GBU*	Still-to-still	Algorithms outperformed humans on frontal face image pairs with "good" and "moderate" viewing conditions and were comparable on "ugly" face image pairs. Face images primarily contained variations in illumination and hairstyle.

*Authors used subsets of the database

Table 1. A comparison of studies in the literature on face recognition performance by humans.

inal face images (99.20% accuracy), (ii) tightly cropped faces without background (97.53% accuracy), and (iii) inverse-cropped images with only the background (94.27% accuracy) [10]. In a follow-on study, Kumar asked MTurk workers whether or not they were familiar with (*i.e*. could identify) either of the individuals in a pair and found that familiarity did not impact face verification accuracy [9]. Until only recently [18, 19], no algorithms had been reported that outperformed these human results.

Video-to-video face recognition studies, compared to still-to-still face recognition, are relatively few with regards to human performance. Motion cues have been shown to improve human accuracy on recognition tasks involving familiar faces, but results have been mixed and inconclusive for unfamiliar face recognition tasks for a summary, see [15, 17]). Table 1 summarizes the related papers that evaluate human accuracy on face recognition tasks.

3. Unconstrained Face Databases

The LFW database [6] is a collection of 13,233 unconstrained face images of 5,749 unique individuals. Each image is the face bounding box output by the Viola-Jones face detector [20] expanded by a factor of 2.2 then rescaled to 250 by 250 pixels. We evaluate performance on the "View 2" protocol (the evaluation set), which consists of 6,000 face image pairs (3,000 genuine and 3,000 impostor pairs).

The YTF database [21] is a collection of 3,425 unconstrained face videos of 1,595 unique individuals. The YTF database was collected by searching YouTube for the names of the individuals in the LFW database. Faces in each video were detected by the Viola-Jones face detector [20], and video clips were only included in the YTF database if faces were detected in 48 or more consecutive frames. Both protocols consist of the same 5,000 video pairs (2,500 genuine and 2,500 impostor pairs). The LFW and YTF databases both detail "restricted" and "unrestricted" experimental protocols; we follow the restricted protocols in this paper.

4. Data Collection via Crowdsourcing

Amazon Mechanical Turk (MTurk) is a website used for "crowdsourcing" (retrieving information) from a large number of human participants ("workers"). A worker on MTurk can do simple Human Intelligence Tasks (HITs) for a "requester." MTurk provides a way for responses on the exact same task to be collected from many different individuals, so a "crowd" is effectively working together to form one final response or result.

We collected human responses via crowdsourcing on MTurk for three face verification studies:

1. Still-to-still on the LFW database,
2. Video-to-video on *original* YTF database videos, and
3. Video-to-video on *cropped* YTF database videos.

See Fig. 2 for the directions, prompt, and choices given to a worker for each pair of face images or videos from the LFW or YTF databases, respectively. The available responses were modeled after [10, 14] and provide a simple measure of a worker's confidence in their decision for each pair of faces. As shown in Fig. 2, we also collected data on familiarity. If a worker recognized either of the two faces presented, they were asked to check the "Familiar?" box and/or fill in the text box with a name or identifying information (*e.g*., the name of a TV show or movie the individual appeared in). The worker received $0.06 for each HIT completed, which consisted of five face image or video pairs.

MTurk allows requesters to impose "qualifications" that a worker must meet prior to completing a requester's HITs. We required workers to have completed at least 100 HITs on MTurk and have an acceptance rate (HITs accepted ÷ total HITs completed) equal to or greater than 85% on all prior HITs. This information is available on MTurk; any worker not meeting these criteria cannot attempt our HITs.

5. Experimental Details

For human performance on LFW, we collected 10 worker submissions for each of the 6,000 face pairs in the LFW protocol as done in [10] and requested the country of origin from each worker. Out of 307 total workers that completed the LFW study, 27.4% were from the United States (USA), 55.1% were from India, and 11.1% left their country of origin blank. Because the majority of the workers appeared to be from only two countries, for the YTF database studies, we collected 40 submissions for each of the 5,000 face pairs in the YTF protocol – 20 responses from USA workers and 20 responses from workers in India.

Many of the YTF videos include background cues or identifying text that could be used by humans to assist in face verification. Thus, we also collected human responses for YTF videos in which the background was blacked out and only the moving face was visible. These *cropped* videos were created by applying a black mask to all pixels not within the bounding box of the detected face. This also promotes a fair comparison of humans and algorithms, as algorithms typically conduct matching on cropped face regions. Note that both the original and cropped videos were converted to videos from the still frames provided by the YTF database, so they do not include any sound that could give humans an advantage.

Responses to HITs from a worker were rejected if they met the following criteria: (i) each of the five face pairs in a single HIT was marked with the same response, (ii) that particular HIT had an accuracy rate of 20% or lower (*i.e.* four or more, of the five, pairs in the HIT were incorrect), and (iii) the worker continued to exhibit behavior (i) and (ii) for five or more consecutive HITs. Table 2 gives statistics on the number of workers, HITs completed, and HITs rejected for the three studies.

After collecting human decisions on the YTF videos, we analyzed the false matches. We found that humans gave non-match verdicts to a number of pairs where the ground truth indicated a match. In a majority of these cases, the crowd's decision was actually correct. We noticed that some of the identities associated with the YTF videos are incorrectly labeled; thus, some genuine/impostor ground truth labels in the experimental protocol are also wrong.[6] To accurately establish a human baseline for video-to-video face verification, we choose to report our results using corrected YTF video pair labels. In total, we found 111 genuine video pairs in the YTF protocol that are actually impostors. See Fig. 3 for examples of incorrect ground truth labels from the YTF database and the confidence scores (similarity) from humans that indicated these pairs were errors.

We compare human performance on LFW to the human performance measured by Kumar *et al*. [10] and the current best algorithms reported on LFW: DeepID [18] and DeepFace [19]. We compare human performance on YTF to the current best algorithm reported on YTF: DeepFace [19]. We further compare human performance to a commercial-off-the-shelf (COTS) face matcher. The same COTS matcher is used for both still-to-still face matching on LFW and video-to-video face matching on YTF. The COTS still image face matcher is applied to faces in videos as outlined in [2]. We apply the COTS face matcher to the cropped and aligned face images (*i.e.* a face track that can be assumed to be a sequence of images of the same person) which are provided by the YTF database [21]. For both the still-to-still and video-to-video studies, we get a single similarity score in the range of 0 to 1 output by the COTS face matcher.

[6] We have released the YTF database label errors to the YTF website.

	LFW	Original YTF	Cropped YTF
Total No. of Workers	307	USA: 431 India: 407	USA: 310 India: 96
Avg. No. of HITs Completed Per Worker	39.1	USA: 46.4 India: 49.1	USA: 64.5 India: 210.5
Total No. of HITs Rejected	0	USA: 14 India: 307	USA: 0 India: 191

Table 2. MTurk data collection statistics for the three studies.

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

Pair	Original Label	Score	Correct Label
(a, d)	Genuine	1.55	Impostor
(a, c)	Genuine	3.75	—
(a, b)	Genuine	1.90	Impostor
(b, d)	Genuine	4.85	—
(b, c)	Genuine	1.50	Impostor
(c, d)	Genuine	1.35	Impostor

Figure 3. Examples of incorrect labels in the YTF database and the average similarity score from 20 MTurk workers for each pair.

6. Experimental Results

We report face verification results for all studies as receiver operating characteristic (ROC) curves, plotting true accept rate (TAR) as a function of false accept rate (FAR). FARs are reported up to 10%, as FARs higher than 10% are generally not useful in operational face recognition scenarios. To render human performance, each individual decision was assigned a value from one to five corresponding to responses of "I am sure that they are not the same" to "I am sure that they are the same." Ten (still-to-still) or twenty (video-to-video) human decisions per face pair were then

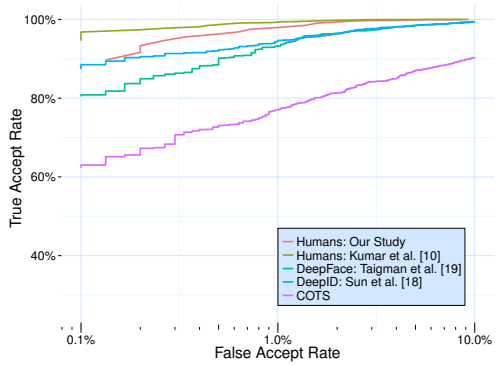


Figure 4. Verification results for still-to-still face pairs from LFW.

(a) Genuine (b) Impostor

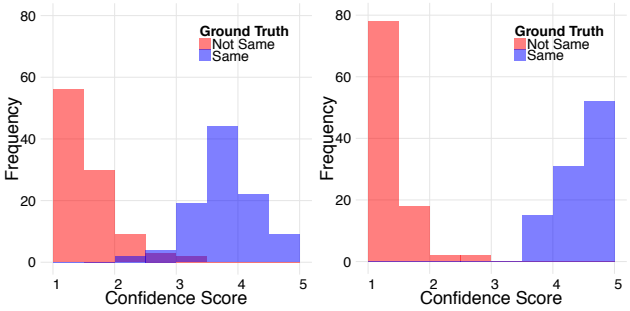
Figure 5. Examples of difficult face pairs from the LFW database.

averaged to compute human "confidence scores." Overall accuracy is also reported by classifying human responses as correct or incorrect. A genuine (impostor) pair is correct if the average score of all workers is greater (less) than three.

6.1. Still-to-Still Face Verification

Figure 4 shows that our result for human performance on the LFW database (98.3% accuracy) is slightly lower, though comparable to Kumar *et al.*'s human performance (99.2% accuracy) [10]. The slight decrease may be due to differences in qualifications of workers between the two studies ([10] had stricter qualification standards, requiring workers to complete at least 1,000 HITs and have an accuracy rate of 95% prior to participating [Kumar, personal communication]). Note also that recently published algorithms, with reported accuracies over 97%, are approaching human performance on LFW [19, 18].

We also measured human performance when presented with two images per subject instead of one. We collected another 20 human responses for each of the 100 lowest (highest) scoring genuine (impostor) face pairs from our first LFW study (Fig. 5). After a worker made a decision based on the "difficult" pair of images, another pair of images of the same two subjects was presented. The worker could then change his/her decision given this new information. Figure 6 shows that given two images per subject, the genuine and impostor distributions of the confidence scores are disjoint; accuracy based on individual responses increased from 80.8% to 91.7% when presented with two images per subject, while accuracy based on crowd confidence scores increased from 95.0% to 100.0%.



(a) (b)

Figure 6. Genuine and impostor score distributions for 200 difficult LFW face pairs given (a) one image per subject and (b) two images per subject.

6.2. Video-to-Video Face Verification

Table 3 summarizes our findings that humans outperform both the COTS face matcher and [19] on YTF (cropped videos) in terms of TARs at low FARs. However, the 91.4% accuracy of same vs. not-same reported by [19] is higher than our measured human accuracies of 89.7% (USA) and 88.6% (India). While all subsequent results leverage our corrected ground truth YTF labels, this table does not, in order to allow for fair comparison against prior publications. We note, however, that [19] also reports results on the corrected YTF labels, achieving an accuracy of 92.5%; our human accuracies on the cropped YTF videos and corrected labels are 91.4% (USA) and 90.0% (India).

Figure 7 breaks down human performance across video type (original vs. cropped) and worker country (USA vs. India). These results are the average ROC curves (and confidence intervals) of 1,000 bootstrap samples, each with 2,389 genuine and 2,611 impostor face pairs. Humans outperform the COTS face matcher in all scenarios at FARs greater than 0.4%. Figure 10 offers examples where humans were able to overcome variations in pose and illumination to correctly identify genuine matches and possibly use gender and ethnicity differences to better recognize impostors. The degradation in performance on cropped videos is consistent with the findings of [10] for LFW. This confirms that humans often use cues other than the cropped face (*i.e.* hair, clothing, background), as shown in Fig. 8.

Method \ FAR	0.4%	1.0%	10.0%	Accuracy
Humans (USA)	71.2	80.6	96.7	89.7
Humans (India)	44.9	63.7	92.4	88.6
COTS	46.3	54.4	81.4	n/a
DeepFace [19]	25.9	54.8	92.0	91.4

Table 3. Comparison of human and algorithm face matching on cropped videos using the original YTF database protocol (*i.e.* no corrected labels), reported as TAR (%) at fixed FARs and accuracy (%) of same vs. not-same decision.

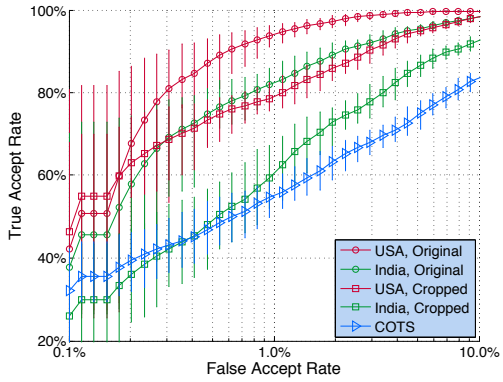


Figure 7. Verification results video-to-video face pairs from YTF.

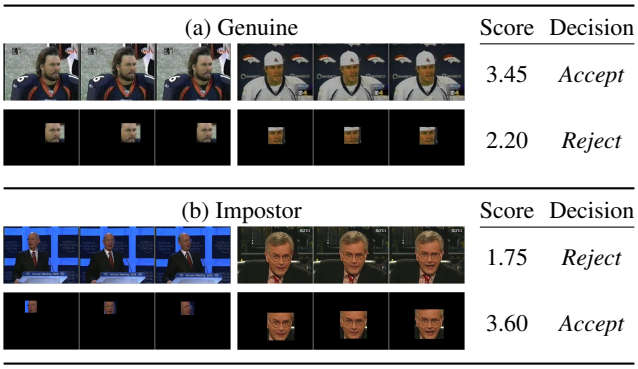


Figure 8. Face pairs where similarity scores given by humans (USA workers) viewing the original YTF videos correctly accepted/rejected at FAR = 1%, while scores given by humans viewing the cropped YTF videos did not.

Workers in the USA consistently outperforming those in India is also significant. This suggests caution in reporting baseline performances from crowdsourcing studies. There are many possible explanations for this phenomenon, including the other-race effect [11], as subjects in the YTF and LFW databases tend to be from Western Europe or the United States. Average accuracies from 100 trials of randomly sampling one unfamiliar response for face pairs for each race demographic are given in Table 4. Race labels were obtained via a separate crowdsourcing study with ten MTurk worker labels per subject (given all LFW images of that subject). There are 4,350 White, 168 Asian, and 217 non-White/Asian race face pairs (remaining 265 face pairs are White/Asian).[7] We observe that workers from USA are more accurate on White than on Asian face pairs, while workers from India perform more consistently on both races. For this reason, we conclude that the other-race effect is present, but to what extent it affects accuracy is hard to determine. Factors such as poorly motivated workers, spammers, and the relatively few number of workers

[7] A White (Asian) face pair contains at least one White (Asian) subject and no Asian (White) subject.

Country	Study	White	Asian	non-w/A
USA	Original	86.0 ± 0.5	81.6 ± 2.2	81.9 ± 1.9
India	Original	83.0 ± 0.5	83.6 ± 2.4	79.5 ± 2.4
USA	Cropped	82.0 ± 0.5	77.5 ± 2.5	73.0 ± 2.3
India	Cropped	78.7 ± 0.6	77.7 ± 2.9	74.4 ± 2.5

Table 4. Accuracies (%) of individual MTurk worker responses for unfamiliar face pairs with respect to race demographics of the two faces in question (*i.e.* White, Asian, non-White/Asian).

Country	Study	unfamiliar	familiar
USA	Original	86.3 ± 0.5	93.5 ± 0.3
India	Original	83.1 ± 0.6	87.8 ± 0.3
USA	Cropped	85.7 ± 1.3	92.6 ± 0.6
India	Cropped	80.4 ± 1.4	82.9 ± 0.4

Table 5. Average ccuracies (%) of individual MTurk worker responses for both unfamiliar and familiar responses.

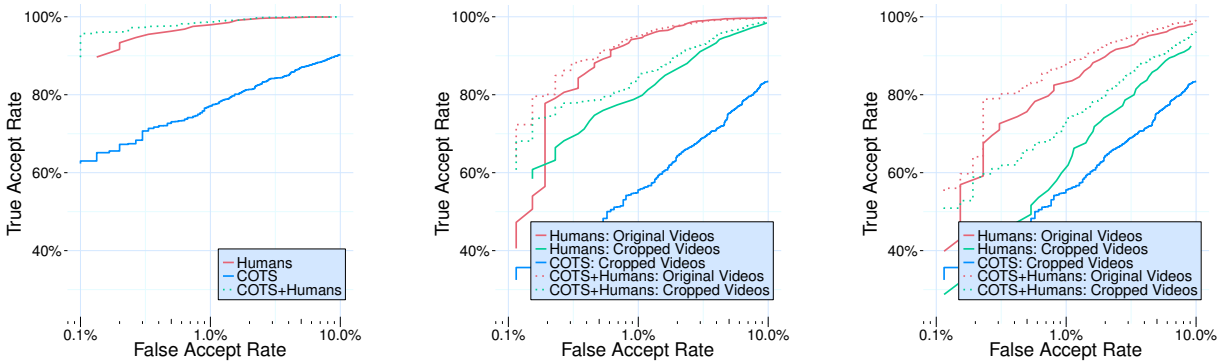
surveyed are also viable explanations for the performance gap between USA and India. Finally, as explored in the following section, familiarity of USA workers to subjects in the YTF database could also contribute to higher accuracy.

6.3. Familiarity

Human performance for face verification tasks has been shown to be better on familiar than on unfamiliar faces [3, 4]. Humans can draw upon different views of a familiar face from their memory to facilitate the face verification task at hand. For example, if a human is presented with a pair of images or videos with limited information (*e.g.* non-frontal view, poor lighting, low-quality), they could draw a connection between the presented image and other images from their memory in order to judge the match. Table 5 indicates that familiarity improves face recognition; individual MTurk worker responses were considerably more accurate on all four YTF studies when the worker reported one or both faces as familiar. These accuracies are averages of 100 trials of randomly sampling a single familiar or unfamiliar response from the set of 3,042 (596) face pairs that contained both familiar and unfamiliar responses for the original (cropped) study. Furthermore, the frequencies with which workers reported familiar faces are 16.4, 11.1, 8.8, and 1.2 percent of the 100,000 responses for USA original, USA cropped, India original, and India cropped studies, respectively. Though it is difficult to prove causality, it is interesting to note that these results correlate strongly with the verification accuracies in Fig. 7 (*i.e.* protocols where workers performed at a higher accuracy also had a greater frequency of familiar faces).

6.4. Fusion of Human and COTS Match Scores

We further evaluate whether human performance can be improved by utilizing complementary information from the COTS face matcher. To combine the human confidence



(a) LFW (b) YTF (USA) (c) YTF (India)

Figure 9. Verification results for fusion of human and COTS match scores for (a) still-to-still study on the LFW database, (b) video-to-video study on YTF database for USA workers, and (c) video-to-video study on YTF database for India workers.

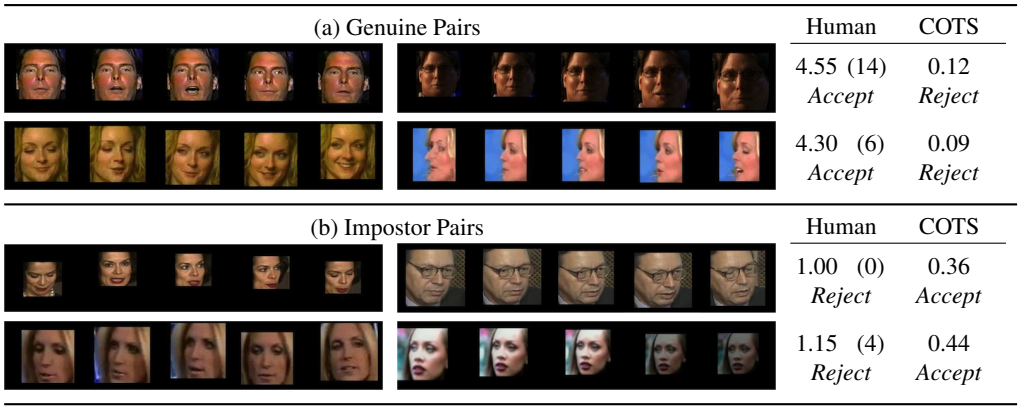


Figure 10. Examples of cropped YTF face pairs where humans (USA workers) were more accurate than COTS. The accept/reject is based on the threshold of Human or COTS similarity scores at FAR = 1%. The number in parentheses represents the number of familiar responses.

scores and the COTS match scores, we first employed a normalization scheme (z-score, tanh, min-max, median). We then applied various fusion rules (min, max, sum, product) to obtain the final fused score for a face pair. We found that z-score normalization with sum fusion performed the best for all studies. Results of fusing human confidence and COTS match scores in this manner are given in Fig. 9. We observe performance improvements with fusion for all five studies presented in this paper.

6.5. Single Human vs. Crowd Performance

With the exception of Tables 4 and 5, all human results presented above refer to the performance of a crowd (*i.e.* average score of 10 or 20 workers). Though commonly done in the literature (particularly for ROC comparison with algorithms), measuring human performance in this manner tends to overestimate "human" (*i.e.* a single person's) accuracy. Table 6 shows the average accuracies of randomly sampling (100 times) a single response per pair in the YTF protocol; the crowd accuracies are higher. As shown in Table 7, the average accuracies of the 20 workers who completed the most HITs are comparable to the accuracies in Table 6 for each YTF study. However, for all studies, the variation in these 20 individual accuracies is quite large, indicating that single human performance greatly depends on the individual.

Country	Study	Rand. Ind. Acc.	Crowd Acc.
USA	Original	87.6 ± 0.4	95.8
India	Original	84.0 ± 0.5	93.7
USA	Cropped	83.2 ± 0.4	91.4
India	Cropped	79.0 ± 0.5	90.0

Table 6. Average accuracy (%) and standard deviation of randomly sampling (100 times) a single response per pair compared with the overall crowd accuracy (%) for each of the YTF studies.

Country	Study	Num. HITs	Accuracy
USA	Original	$1,658 \pm 434$	89.1 ± 9.0
India	Original	$1,565 \pm 609$	84.7 ± 6.3
USA	Cropped	$1,900 \pm 367$	83.4 ± 4.9
India	Cropped	$3,621 \pm 1,005$	78.9 ± 10.1

Table 7. Average number of HITs completed and average accuracies (%) of the 20 individual MTurk workers who completed the most HITs for each of the YTF studies.

7. Conclusions

This paper presented human performance on unconstrained still-to-still and video-to-video face matching scenarios. The two face databases used, LFW and YTF, are publicly available and commonly used to evaluate the performance of face recognition algorithms on unconstrained faces in still images and videos. We compared our measured human accuracies (obtained via MTurk) to the performance of published algorithms and a COTS face matcher. Some key findings of this paper are: (i) Humans perform better (in terms of TARs at low FARs) than state-of-the-art face recognition systems, including a COTS face matcher, on still-to-still and video-to-video matching of unconstrained faces; machine accuracies on same vs. not-same face classification are now comparable to humans. (ii) Humans make use of contextual information, or soft biometrics (*e.g.* ethnicity, gender, hairline, clothing, background). (iii) Human accuracy can depend on the demographics of the crowd workers; human accuracy improves when workers are familiar with one or both of the faces in question. (iv) Fusion of human confidence and COTS match scores improves performance for still-to-still and video-to-video face matching, implying that human and COTS decisions offer some complementary information for unconstrained faces. (v) Crowd performance tends to overestimate the performance of a single human, and single human performance greatly depends on the individual.

Additionally, by analyzing the (supposed) matching errors made by humans, we discovered ground truth labeling errors in the YTF database. That is, when looking closely at pairs with low genuine scores (or high impostor scores), we discovered that humans were actually correct, while the label of genuine/impostor pair was wrong. We believe these labeling errors are due to the nature of how the YTF database was compiled. The issue with searching the web for videos tagged with a specific name is that each face track extracted from a video needs to be verified that it actually corresponds to the person of interest. We suggest that before publicly releasing a database, crowdsourcing should be used to help verify the ground truth labels of the database.

References

[1] A. Adler and M. E. Schuckers. Comparing human and automatic face recognition performance. *IEEE Trans. on SMC - Part B*, 37(5):1247–1255, Oct. 2007.

[2] L. Best-Rowden, B. Klare, J. Klontz, and A. K. Jain. Video-to-video face matching: Establishing a baseline for unconstrained face recognition. In *Proc. BTAS*, 2013.

[3] V. Bruce, Z. Henderson, C. Newman, and A. M. Burton. Matching identities of familiar and unfamiliar faces caught on cctv images. *Journal of Experimental Psychology: Applied*, 7(3):207–218, Sep. 2001.

[4] A. M. Burton, S. Wilson, M. Cowan, and V. Bruce. Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, 10(3):243–248, May 1999.

[5] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *Proc. ECCV*, 2012.

[6] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Report 07-49, Univ. of Mass., Amherst, Oct. 2007.

[7] A. K. Jain, B. Klare, and U. Park. Face matching and retrieval in forensics applications. *IEEE Multimedia*, 19(1):20–28, Jan. 2012.

[8] J. C. Klontz and A. K. Jain. A case study on unconstrained facial recognition using the boston marathon bombing suspects. Tech. Report MSU-CSE-13-4, Michigan State Univ., East Lansing, MI, USA, May 2013.

[9] N. Kumar. *Describable Visual Attributes for Face Images*. PhD thesis, Columbia Univ., New York, 2011.

[10] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Trans. PAMI*, 33(10):1962–1977, Oct. 2011.

[11] C. A. Meissner and J. C. Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3–35, Mar. 2001.

[12] National Institute of Standards and Technology (NIST). Face homepage. `http://face.nist.gov`, Jun. 2013.

[13] 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. on Applied Perception*, 9(4):16:1–16:13, Oct. 2012.

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

[15] A. J. O'Toole, D. A. Roark, and H. Abdi. Recognizing moving faces: A psychological and neural synthesis. *TRENDS in Cognitive Sciences*, 6(6):261–266, Jun. 2002.

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

[17] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proc. IEEE*, 94(11):1948–1962, Nov. 2006.

[18] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. CVPR*, 2014.

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

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

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