

# MegaFace: A Million Faces for Recognition at Scale

D. Miller E. Brossard S. Seitz I. Kemelmacher-Shlizerman  
 Dept. of Computer Science and Engineering  
 University of Washington

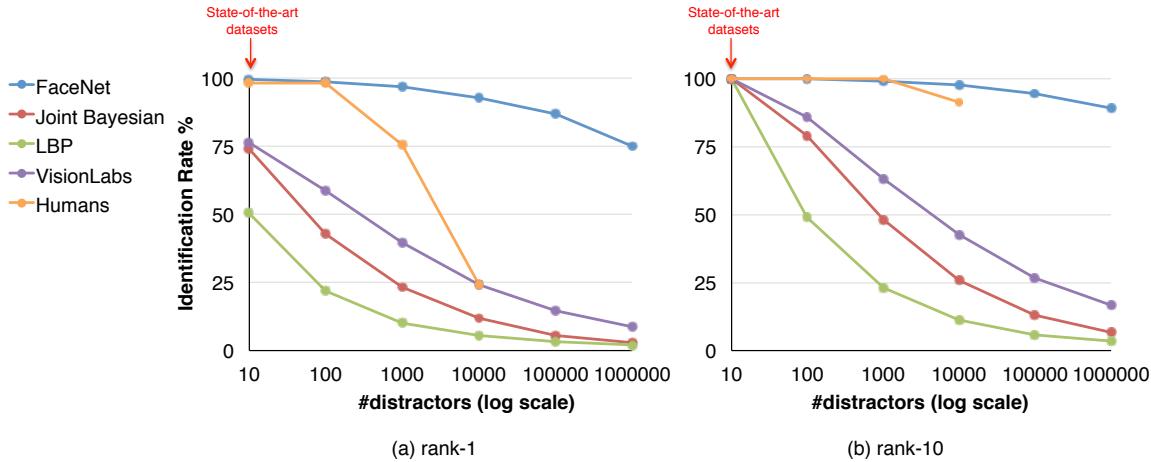


Figure 1: We evaluate how recognition performs with increasing numbers of faces in the database: (a) shows rank-1 identification rates, and (b) rank-10. Recognition rates drop once the number of distractors increases. We also present first large-scale human recognition results (up to 10K distractors). Interestingly, Google’s deep learning based FaceNet is more robust at scale than humans. See <http://megaface.cs.washington.edu> to participate in the challenge.

## Abstract

Recent face recognition experiments on the LFW [13] benchmark show that face recognition is performing stunningly well, surpassing human recognition rates. In this paper, we study face recognition at scale. Specifically, we have collected from Flickr a **Million** faces and evaluated state of the art face recognition algorithms on this dataset. We found that the performance of algorithms varies—while all perform great on LFW, once evaluated at scale recognition rates drop drastically for most algorithms. Interestingly, deep learning based approach by [23] performs much better, but still gets less robust at scale. We consider both verification and identification problems, and evaluate how pose affects recognition at scale. Moreover, we ran an extensive human study on Mechanical Turk to evaluate human recognition at scale, and report results. All the photos are creative commons photos and are released for research and further experiments on <http://megaface.cs.washington.edu>.

## 1. Introduction

Face recognition has seen major breakthroughs in the last couple of years, with new results by multiple groups [23, 27, 25] **surpassing human performance** on the leading Labeled Faces in the Wild (LFW) benchmark [13] and achieving near perfect results.

Is face recognition solved? Many applications require accurate identification at *planetary scale*, i.e., finding the best matching face in a database of billions of people. This is truly like finding a needle in a haystack. Face recognition algorithms did not deliver when the police was searching for the suspect of the Boston marathon bombing [14]. Similarly, do you believe that current cell-phone face unlocking programs will protect you against anyone on the planet who might find your lost phone? These and other face recognition applications require finding the true positive match(es) with negligible false positives.

In this paper, we introduce the *Megaface* dataset and benchmark for large scale face recognition. The goal of this dataset is to evaluate the performance of current face recognition algorithms with up to a million *distractors*, i.e.,

up to a million people who are not in the test set. Our key objectives for this dataset are that it should 1) contain photos “in the wild”, i.e., with unconstrained pose, expression, lighting, and exposure, 2) contain regular people, not easily recognizable celebrities, 3) be broad rather than deep, i.e., contain many different people rather than many photos of a small number of people, and 4) be publicly available, to enable distribution within the research community. Whereas previous face benchmarks have relied on celebrity photos, passport photos, or mugshots, our objectives require a different approach. Instead, we leverage the recently released database of Flickr Creative Commons photos [29], from which we extracted 1 million faces (randomly sampling the full 100M photo collection). We intend to release even larger datasets (from the full 100M collection) in the future, setting aside training and testing sets to ensure an even playing field.

Based on this new benchmark, we address fundamental questions and introduce the following key findings:

- **How well do current face recognition algorithms scale? Key finding:** While performance **drops by 70%** for most algorithms, Google’s [23] deep-learning based FaceNet achieves 75% identification rate even with a million distractors (Fig. 1).
- **How well does human face recognition scale?** Even devising a practical human face identification experiment (requiring people to sort lists containing thousands of faces) is challenging. We performed the first large scale human face identification experiment, leveraging a crowd of Mechanical Workers to collectively sort the best matches from each probe image against a database containing one true match and ten thousand distractors. Humans’ rank-1 identification rate is 23.9% with 10K distractors, 91.13% at rank-10.
- **How does pose affect recognition performance?** Somewhat surprisingly, recognition rates **drop** when comparing frontal-to-frontal images, compared to the task of comparing faces when the pose is not controlled.

While this paper benchmarks only a sparse sampling of face recognition algorithms, we believe that it’s a reasonable initial sampling, as it includes the current top performer on LFW [23], another commercial LFW top-performer (VisionLabs), and two baseline algorithms (Joint Bayesian and LBP) that are popular in the academic community. Furthermore, if the paper is accepted to ICCV, we will maintain and update this benchmark online and solicit contributions from the other top performers (e.g., we are in contact with Facebook and hope to include DeepFace [27] and others before the paper goes out to press).

The remainder of this paper is organized as follows. We first describe related face datasets and benchmarking efforts, and the details of our dataset. We then describe our evaluation methodology, our efforts at evaluating human performance at large scale, and our benchmark of face recognition algorithms. We then analyze the effects of pose and recognition accuracy, and conclude the paper.

## 2. Related Work

Early work in face recognition focused on controlled datasets where subset of lighting, pose, or facial expression was kept fixed, e.g., [8, 9]. With the advance of algorithms, the focus moved to unconstrained scenarios with a number of important benchmarks appearing, e.g., FRGC, Caltech Faces, and many more (see [13] Fig 3. for a list of all the databases), as well as, thorough evaluations were made [11, 34]. A big challenge, however, was to collect photos of large number of individuals.

In 2007, Huang et al. [13] created the benchmark Labeled Faces in the Wild (LFW). The LFW database includes 13K photos of 5K different people. It was collected by running Viola-Jones face detection [30] on Yahoo News photos. LFW captures celebrities photographed under completely unconstrained conditions (arbitrary lighting, pose, and expression) and it turned out to be an amazing resource for face analysis community. Since 2007, a number of databases appeared that include larger numbers of photos per person (LFW has 1620 people with more than 2 photos), video information, and even 3D information [15, 4, 33, 31, 5, 18]. However, LFW is still the leading benchmark on which all state of the art recognition methods are evaluated and compared. Indeed, just in the last few months a number of methods [23, 26, 25, 27, 28] reported recognition rates above 99%+ [12] (better than human recognition rates estimated on the same dataset by [16]). The perfect recognition rate on LFW is 99.9% (it is not 100% since there are 5 pairs of photos that are mislabeled).

Our paper is about taking the face recognition task to the next level and revealing the challenges that appear once recognition is done on large scale (many orders of magnitude larger than current evaluation). Specifically, when the number of photos in the database includes 1 Million faces versus tens of thousands (Fig 2). Companies like Facebook and Google have access to extremely large numbers of photos, e.g., Facebook has trained on 4 Million photos of 4K people [27], Google [23] trained on 200 Million photos of 8 Million people. These datasets, however, are not available to the public and were used only for training and not testing. Large scale evaluations were performed only on controlled datasets (visa photographs, mugshots, lab captured photos) by NIST [11], and report recognition results of 90% on 1.6 million people. We show that recognition rates on uncon-

| Dataset                  | Available     | #Photos and #people          |
|--------------------------|---------------|------------------------------|
| LFW                      | Public        | 13K of 5K people             |
| CelebFaces 2014          | Private       | 202K of 10K people           |
| CASIA-WebFace 2014       | Public        | 500K of 10K people           |
| FaceScrub 2014           | Public        | 100K of 500 people           |
| YouTube Faces            | Public        | 3425 videos of 1595 people   |
| DeepFace (Facebook) 2014 | Private       | 4.4 Million of 4K people     |
| FaceNet (Google) 2015    | Private       | 100-200 Million of 8M people |
| <b>MegaFace</b>          | <b>Public</b> | <b>1 Million</b>             |

Figure 2: Representative sample of face recognition datasets that were created in the recent years (in addition to LFW). All the public datasets are small scale, and all the large scale datasets are mainly used for training rather than testing and are not publicly available. MegaFace (this paper) is the first large scale unconstrained dataset. It is collected from Flickr and will be available publicly.

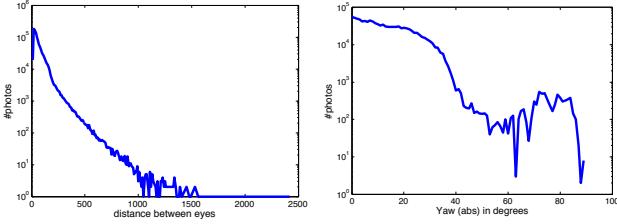


Figure 4: Number of photos per resolution (left) and distribution of poses in MegaFace. Resolution is measured by the distance between the eyes (x-axis). For comparison LFW’s distance is 40 while our MegaFace dataset has a wider distribution since Flickr photos are typically high resolution photos taken with DSLR cameras.

strained data are much lower. [19] experiment with large scale identification assuming there is more than one input photo per person. Recently, several papers reported identification results on the LFW dataset [3, 28, 26, 25].

### 3. The 1 Million Faces Dataset

Our goal is to evaluate how the size of a dataset affects recognition rates. Specifically we’re interested in varying the number of “distractors”, i.e., people that are not in the test set, and evaluating performance of current face recognition algorithms. For example, consider the scenario of identifying a person from a single photo, by comparing that image (the “probe”) to a database of billions of other people. This would require several orders of magnitude more comparisons than LFW, which only involves a few thousand people. We decided that a good starting point would be to evaluate recognition with a million people. Specifically, we

had the following challenges:

- unconstrained in-the-wild photos, i.e., any lighting, pose, expression, age, and resolution
- photos of regular people, i.e., not celebrities
- 1 million photos
- publicly available to enable further experiments by other researchers
- broad rather than deep dataset, i.e., large number of identities versus many photos of a small set of people.

While there are massive amounts of photos on the Internet, satisfying the above requirements turned out to be not trivial. Indeed, previous works that released public datasets focused on photos of celebrities or constrained environments, and collected photos of at most 10K different people. Governmental datasets, e.g., drivers licenses, mugshots, etc. are not available for research and typically captured under constrained conditions. Datasets that are collected by Google, Facebook, Face++ and others, similarly are not available for research.

This year, however, Yahoo released their 100 Million Flickr creative commons dataset [29], which turned out to be a perfect opportunity to test face recognition at scale. We randomly sampled images from this dataset, detecting faces in the photos until we collected a million faces. These photos are uploaded by Flickr users under the creative commons license and thus the chance of celebrities occurring is exceedingly small (photos of celebrities are typically not available under the creative commons license). We do not guarantee 1 million unique faces (different people) however we optimized for unique users and large fraction are group photos which ensure that the people in the photo are different. Specifically, we have downloaded photos from 178,452 different users. Given that people appearing in the same group photo are different, we assembled total of 690,572 unique faces (counting a single photo per user). The rest of the faces may be unique as well, however, we cannot guarantee this. This is the largest set of unique people ever assembled.

**Data collection protocol.** We began by estimating that there are 500K userids in the Flickr set. Our algorithm for detecting and downloading faces was as follows. For each userid we consider previously unseen photo (based on name, quality, and resolution) and test whether it is possible to detect faces in this photo. If the photo does not have faces we continue to the next one in that user’s set, otherwise we detect all the faces in the photo (many of the photos are group photos). Once a photo with faces is found for a particular user we continue to the next user. We terminate the process once we’ve reached a million faces. We downloaded the highest resolution available per photo. The faces

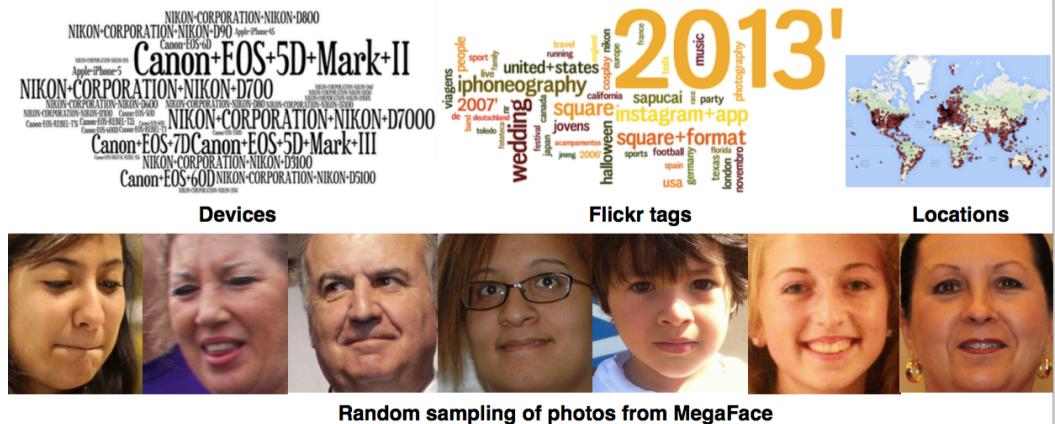


Figure 3: The MegaFace dataset: distributions of devices, Flickr tags, and location. We also show a random sample of the photos in the dataset. All the 1 Million photos in the dataset are creative commons photos and will be released for research.

are detected using the HeadHunter<sup>1</sup> algorithm by Mathias et al. [17], which reported state of the art results in face detection (see exact detection rates and comparisons to others methods in their paper). We also found that it is robust to extreme face poses. For each face we save the detected face (with the face taking about 75% of the photo). We further estimate 49 fiducial points, and yaw and pitch angles, all calculated by the IntraFace<sup>2</sup> landmark model [32].

**Data statistics.** In Figure 3, we present statistics of the dataset. Specifically, we plot distributions of resolution, location, pose, devices, tags, and user ids. The dataset has good distribution of locations, most of the photos were captured by DSLR cameras, tags include words from ‘instagram’ to ‘wedding’ which suggests a range of photos from selfies to high quality portraits (large amount of the photos came with a tag ‘2013’ due to nature of the dataset—recently uploaded photos). We also report the distribution of yaw angle and resolution of the face via the distance in pixels between the eyes. More than 50% (514K) of the photos in MegaFace have resolution more than 40 pixels inter-ocular distance (which is the resolution of LFW photos). In addition, we inspected by running stricter face detectors on the automatically downloaded photos, as well as, manually inspected 1500 random photos, and found that 6% of the photos are very blurry, non faces, or very low resolution.

#### **4. Evaluation Methodology**

Given the dataset, we ran several experiments to examine two classic recognition scenarios: identification and verification. In addition, we ran a large scale human performance evaluation on the identification task. Below we describe the

methodology and our test set.

**Test set.** Our Flickr dataset is used to create a large number of distractors. For testing known identities we use a subset of the FaceScrub dataset that was created in 2014 by Ng and Winkler [18]. FaceScrub includes 100K photos of 530 celebrities. It can be freely downloaded from <http://vintage.winklerbros.net/facescrub.html>. We chose to use this dataset as our test set (rather than LFW) since it has a similar number of male and female photos (55,742 photos of 265 males and 52,076 photos of 265 females) and a large variation across photos of the same individual. Ensuring variation within individual’s photo collection is important to remove possible bias, e.g., due to backgrounds and hair style [16], that may occur in LFW. We randomly selected a subset of FaceScrub to create a set which is comparable to LFW in size but has more variation across identities. We randomly selected 80 identities with that had more than 50 images each, for a total of 4000 faces.

**Identification and Verification.** Identification means that given a photo of a person as input and assuming there is a photo of this person in the database together with many other people, the algorithm should match the two photos of the same person. Verification means that given a pair of photos, the algorithm should output whether the person in the two photos is the same or not. Until now mostly verification was at the focus of face recognition research, and tested by the LFW benchmark [13]. Recently, [3, 28, 26, 25] performed also identification experiments on LFW. In this paper, we explore both scenarios but with very large number of distractors.

Specifically, each person was enrolled to the database

with a single photo, i.e., the database included a million Flickr photos and photo per test identity. Then for each person we used as test photo every photo from the collection except for the database one. We then averaged the results over the different photos. We report the results as Cumulative Match Characteristics curves. This curve shows the probability that the correct gallery image will be chosen for a random probe by rank-K in a list of results. To evaluate verification we computed all pairs between the probe database (FaceScrub) and the distractor database (Flickr). This means that our verification experiments has in total 4 billion negative pairs. We report verification results with ROC curves; this explores the tradeoff between falsely accepting non-match pairs and falsely rejecting match pairs.

**Recognition Methods.** We have selected to experiment with four recognition algorithms that represent very different types of techniques.

- **Basic LBP comparison.** We have implemented a comparison based on Local Binary Pattern (LBP) descriptors [2]. This approach achieves 70% recognition rates on LFW, and uses no training.
- **Joint Bayesian.** The Joint Bayesian model represents each face as the sum of two Gaussian variables  $x = \mu + \epsilon$  where  $\mu$  represents identity and  $\epsilon$  represents inter-personal variation. To determine whether two faces,  $x_1$  and  $x_2$  belong to the same identity, we calculate  $P(x_1, x_2|H_1)$  and  $P(x_1, x_2|H_2)$  where  $H_1$  is the hypothesis that the two faces are the same and  $H_2$  is the hypothesis that the two faces are different. These distributions can also be written as normal distributions, which allows for efficient inference via a log-likelihood test. This algorithm, trained on [33] achieves 89% on LFW.
- **Commercial software by VisionLabs.** VisionLabs <http://www.visionlabs.ru/> achieved 93% recognition rates on LFW, and is trained on outside data.
- **Google’s FaceNet.** Google FaceNet [23] is the most recent and highest performing of several Deep Learning algorithms applied to the LFW benchmark. Unlike DeepFace and DeepID which have a bottleneck layer and are optimized by minimizing cross-entropy, FaceNet learns an embedding such that the extracted features are directly comparable using the Euclidean distance. It is trained on 200 Million photos of 8 Million people, and achieves 99.6% on LFW.

## 5. Human Performance

While a lot of effort goes into developing automated face recognition algorithms, still the human visual system seems to perform better especially at very low false accept rate. It

is therefore very interesting to estimate the human recognition rate on the same sets of photos on which algorithms operate. Verification rates of humans were previously estimated on controlled data, i.e., photos taken in laboratory condition [21, 20, 22, 1], and more recently on unconstrained photo collections: [16] evaluated verification rates on the LFW dataset, and [3, 6] estimated verification rates on videos. Human studies made on unconstrained photos, e.g., [16], fused human judgments by averaging ratings over participants, which helped remove outliers. Until recently, none of the algorithms was able to outperform humans on LFW. Moreover, all the previous human experiments were done on small scale and did not evaluate identification. It is of great interest, however, to discover how humans perform at *scale* to provide a lower bound for machine performance.

One of the key contributions of this paper is an extremely large (about 4 million pairs of faces) human study on Mechanical Turk that evaluates human performance on unconstrained photos (Flickr), and specifically targets identification rather than verification.

We have performed the following experiment on Mechanical Turk. Since all the identities in the FaceScrub dataset (our test set) are celebrities, human recognition rates may be biased due to familiarity with the person [24]. We therefore sorted all the names in FaceScrub according to the number of results that Google image search returns per person as a measure of popularity. We then chose 50 most popular people, and 50 least popular people as our human experiment test set. Each person had 100 photos, we randomly selected one photo as the probe image, and used the rest 99 as gallery images. We then produced 99 positive pairs per person. For the distractor set, for each input photo we randomly selected 10K photos from our MegaFace dataset, and produced 10K pairs of probe with each of the distractors. This results in total of  $100 \times (99 + 10K)$  pairs. Since the number of positive pairs in this setting is very low, we introduced additional positive pairs by randomly pairing gallery images that are not the probe. This is to remove possible bias in human rating, i.e., if most pairs are negative people may miss the positive ones. We presented to turkers 10 pairs per page and asked to click on all the pairs that contained the same person. We paid 1 cent per page of 10 pairs.

Once this experiment was done we collected the pairs that received 1 click or more, and created a sorting experiment. We selected only the pairs that include the probe photos, and created a set of possible matches per probe. We generated triples of probe, and two matches, presented 10 triples in each page and asked which one of the matches is the person in the probe. Generally, to get a full ranking of all images, the number of possible triples per probe is  $n^2$  where  $n$  is the number of matches from round 1. For efficiency (and less cost) only determined the position of each gallery photo relative to the distractor images. That is, our

| Identification |        |         |
|----------------|--------|---------|
|                | Rank-1 | Rank-10 |
| All            | 23.9   | 91.13   |
| Males          | 23.35  | 89.98   |
| Females        | 24.01  | 92.5    |
| Less Popular   | 22.7   | 90.9    |
| More Popular   | 25.1   | 91.3    |

| Verification |                          |                          |
|--------------|--------------------------|--------------------------|
|              | TAR @ $2 \times 10^{-3}$ | TAR @ $5 \times 10^{-2}$ |
| All          | 41.6                     | 76.5                     |
| Males        | 43.7                     | 79.0                     |
| Females      | 39.4                     | 73.9                     |
| Less Popular | 39.4                     | 74.7                     |
| More Popular | 43.6                     | 78.2                     |

Figure 5: Human recognition rates (verification and identification). Our experiments also show that humans perform better on more popular people and are better at the verification task when comparing males.

experiment determined the number of distractors that would be ranked above and below each gallery image, but not the ordering within those groups. On every pair/triple of photos in both experiments worked 3 different people. We paid 7 cents for each page of 10 triples. The total cost of this experiment was \$10,000.

## 6. Benchmarking Recognition at Scale

This section describes the verification and identification experiments we performed with the MegaFace dataset as our distractor set. In addition, we describe the effect of pose variation on recognition at scale. We have experimented with four automatic face recognition algorithms: LBP and Joint Bayes were implemented by us, VisionLabs has provided their software for our experiments, and FaceNet algorithm was ran by the authors on our data. Prior to our experiments, we have verified for all methods that they achieve their reported results [13] on the LFW dataset. Similarly, we repeated the human study of [16] on LFW using our Mechanical Turk interface to ensure that our results are valid (Fig. 2 in the sup. material).

**Verification.** Fig. 6 shows results of a verification experiment of two algorithms (FaceNet and Joint Bayesian) with various numbers of distractors going from 10 to 1 million. We make two observations:

- Verification does not change much at scale, particularly when we consider false accept rates as in LFW.

Results at LFW are typically reported at equal error rate which implies false accept rate of 1%-5% for top algorithms. We believe the reason that rates stay constant at scale is because given a probe photo, if the face has 100 other faces that can be matched wrongly in a small dataset, e.g., thousand faces, assuming uniform distribution of the data, the rate will stay the same, and so in a dataset of a million faces one can expect to find 10,000 matches at the same false accept rate.

- Striving to perform well at low false accept rate is important with large datasets. Even though the chance of a false accept on the small benchmark is acceptable, it does not scale to even moderately sized galleries.

**Identification.** The relation between the identification and verification protocols was studied by Grother and Phillips [10] and DeCann and Ross [7]. Only recently, however, identification results have started to appear on the LFW dataset. Here we evaluate identification with large number of distractors. In Figs. 1 and 7 we show the performance of the four algorithms with respect to different ranks, i.e., rank-1 means that the correct match got the best score from the whole database, rank-10 that the correct match is in the first 10 matches, etc. Fig 1 shows that rates drop significantly at scale for everyone that we tested except for FaceNet which exhibits a relatively small decrease. Fig. 7 shows the behavior at higher ranks.

**Pose.** In general, frontal faces are considered to be easier to match since features are directly comparable and alignment is well understood. However, when we removed all non-frontal images from our Flickr dataset and repeated our experiment, the results were surprising: rates dropped (Fig. 9). We have created a distractor set of only frontal photos ( $\text{yaw} < 2$  degrees), and only non-frontal ( $\text{yaw} > 15$  degrees). We tested the Joint Bayesian, and FaceNet algorithm with this set. We got that identification rates are lower by about 5% in case of a frontal distractor set for the Joint Bayesian algorithm. FaceNet that performs better than JB on the full set, does not exhibit significant difference between frontal and non frontal distractors.

A number of factors could explain this. Our primary hypothesis is that our probe dataset is biased towards frontal images, and are thus less likely to match to non-frontal images in the distractor database. Another plausible explanation of this effect is that algorithms with limited learning capacity must trade performance for a given pose for generalization across pose; i.e., an algorithm that is able to match across pose might perform worse at frontal recognition.

We have also experimented with difference in pose within the pair of photos. Fig. 8 evaluates error in recognition with respect to difference in yaw between the probe

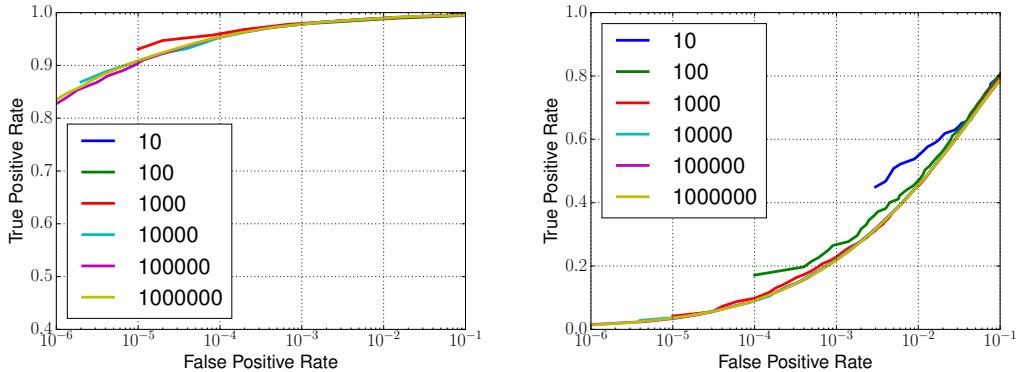


Figure 6: Verification performance with 10, 100, 1K, 10K, 100K, and 1 Million distractors, for two methods: FaceNet (left) and Joint Bayesian (right). Verification rates are stable with different database sizes.

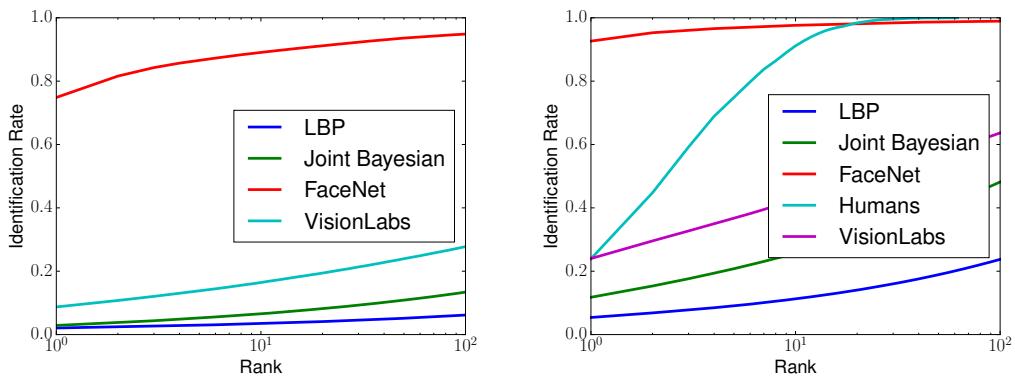


Figure 7: Identification performance for all methods with 1M distractors (left) in the database, and 10K distractors (right) which also includes human performance.

and gallery. We can see that larger pose difference implies larger errors. The results are normalized by the total number of pairs for each pose difference.

## 7. Discussion

Ultimately, a face recognition algorithm should be able to perform even with billions of people in the dataset. While testing with billions is still challenging we have done the first step and created a database of a million faces. This dataset will be available to researchers and we hope that more methods will be tested and improved using the provided data.

In the future, we intend to release all the detected faces from the 100M Flickr dataset. While companies like Google and Facebook have a head start due to availability of enormous amounts of data, we're interested to provide an even playing field for evaluation and training of algorithms at scale. A big challenge is, can one come up with high identification rates by training on our Flickr data. Our

dataset will be separated to testing and training sets for fair evaluation and training. This playing field point is particularly important, as the Google + Facebook methods are trained on orders of magnitude more imagery than the others.

We will maintain and update this benchmark online and solicit contributions from the other top performers (e.g., we are in contact with other companies and hope to include additional results before the paper goes out to press). Our first challenge is to test identification and verification with 1 million distractors, while the FaceScrub dataset is used as the test set. We plan to keep creating more challenges in the future. One challenge will be to create a test set from Flickr photos. An interesting question is whether results will change if test set does not include celebrity photos but regular Flickr photos. Of course dataset bias will be taken into account. We will also test different sizes of test set going from 80 identities (current challenge) to hundreds of thousands. Another challenge will be to allow multiple pho-

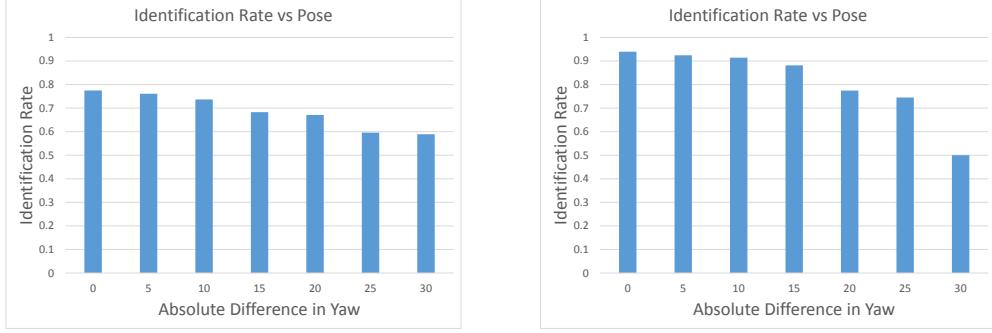


Figure 8: Identification performance of FaceNet (left plot) and humans (right plot) as a function of pose. FaceNet (left): we can see that performance decreases with pose difference (between the probe and gallery) even with the best method. This indicates that cross-pose matching is still an open problem. Humans (right): similarly to automatic methods humans perform worse with difference in pose.

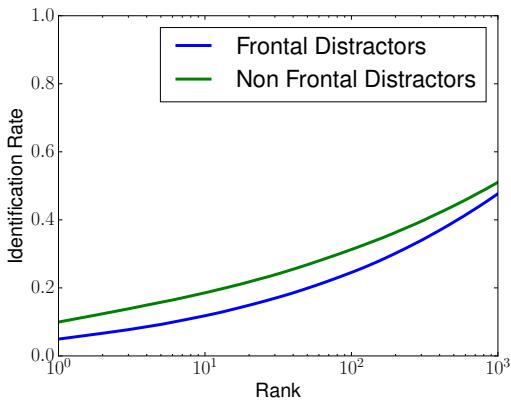


Figure 9: Cumulative Match Characteristics of the Joint Bayesian method with 100K distractors in the database. Recognition rates are lower in case the distractors appear in frontal pose ( $\text{yaw} < 2$  degrees).

tos rather than a single photo per person for identification. Finally, the significant number of high resolution faces in our Flickr database will also allow to explore the issue of resolution in more depth. Resolution is mostly untouched topic in face recognition literature, mostly because public data was not available.

## References

- [1] A. Adler and M. E. Schuckers. Comparing human and automatic face recognition performance. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 37(5):1248–1255, 2007. 5
- [2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition.
- [3] L. Best-Rowden, S. Bisht, J. C. Klontz, and A. K. Jain. Unconstrained face recognition: Establishing baseline human performance via crowdsourcing. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–8. IEEE, 2014. 3, 4, 5
- [4] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, et al. 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. IEEE, 2013. 2
- [5] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Computer Vision–ECCV 2012*, pages 566–579. Springer, 2012. 2
- [6] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *Computer Vision–ECCV 2012*, pages 766–779. Springer, 2012. 5
- [7] B. DeCann and A. Ross. Can a poor verification system be a good identification system? a preliminary study. In *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*, pages 31–36. IEEE, 2012. 6
- [8] A. Georghiades. Yale face database. *Center for computational Vision and Control at Yale University, http://cvc.yale.edu/projects/yalefaces/yalefa*, 1997. 2
- [9] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 2
- [10] P. Grother and P. J. Phillips. Models of large population recognition performance. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–68. IEEE, 2004. 6
- [11] P. J. Grother, G. W. Quinn, and P. J. Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST interagency report*, 7709:106, 2010. 2

*Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006. 5

- [12] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. *arXiv preprint arXiv:1504.02351*, 2015. [2](#)
- [13] 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, Technical Report 07-49, University of Massachusetts, Amherst, 2007. [1](#), [2](#), [4](#), [6](#)
- [14] J. C. Klontz and A. K. Jain. A case study on unconstrained facial recognition using the boston marathon bombings suspects. *Michigan State University, Tech. Rep.*, 119:120, 2013. [1](#)
- [15] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009. [2](#)
- [16] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):1962–1977, 2011. [2](#), [4](#), [5](#), [6](#)
- [17] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014. [4](#)
- [18] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. *people*, 265(265):530. [2](#), [4](#)
- [19] E. G. Ortiz and B. C. Becker. Face recognition for web-scale datasets. *Computer Vision and Image Understanding*, 118:153–170, 2014. [2](#)
- [20] A. J. O’Toole, X. An, J. Dunlop, V. Natu, and P. J. Phillips. Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception (TAP)*, 9(4):16, 2012. [5](#)
- [21] A. J. O’Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Peñard, and H. Abdi. Face recognition algorithms surpass humans matching faces over changes in illumination. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(9):1642–1646, 2007. [5](#)
- [22] 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 experimental results. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):831–846, 2010. [5](#)
- [23] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*, 2015. [1](#), [2](#), [5](#)
- [24] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006. [5](#)
- [25] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. [1](#), [2](#), [3](#), [4](#)
- [26] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *arXiv preprint arXiv:1412.1265*, 2014. [2](#), [3](#), [4](#)
- [27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. [1](#), [2](#)
- [28] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. *arXiv preprint arXiv:1406.5266*, 2014. [2](#), [3](#), [4](#)
- [29] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015. [2](#), [3](#)
- [30] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. [2](#)
- [31] 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. IEEE, 2011. [2](#)
- [32] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013. [4](#)
- [33] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. [2](#), [5](#)
- [34] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003. [2](#)

## 8. Supplementary material

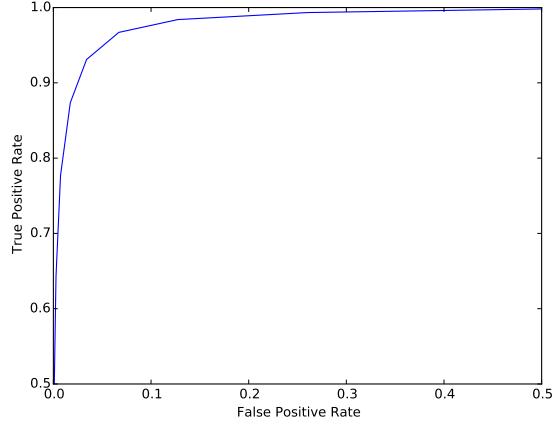


Figure 10: In order to verify that our interface did not affect the performance of Human Recognition, we repeated the LFW experiment by Kumar et al. with our interface – that is, for each of the 6000 pairs in LFW we asked 10 workers to decide whether each pair was same / not same. The rates are slightly lower than Kumar’s et al. results (0.9576 vs 0.9753), but still very good. We suspect that the difference is since we asked only same/not same question (binary) while Kumar et al. asked to rank on a scale of 0 to 5.

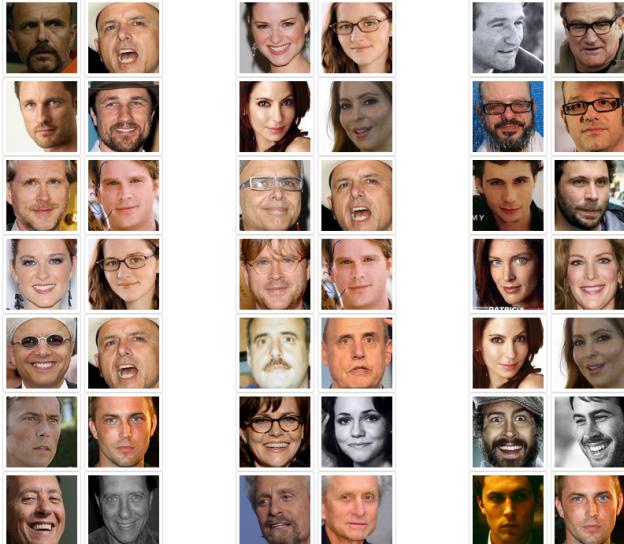


Figure 11: Random pairs of probe/gallery images that were matched correctly by FaceNet, but not by Humans. [Click here to view the full set of faces.](#)



Figure 12: Random pairs of probe/gallery images that were matched correctly by Humans, but not by FaceNet. [Click here to view the full set of faces.](#)

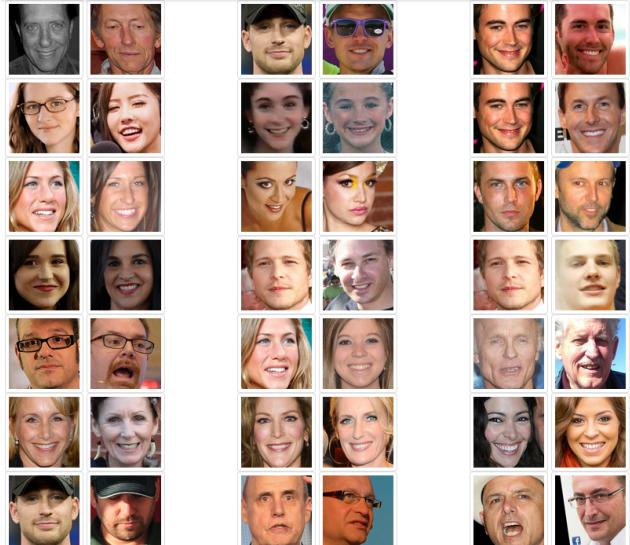


Figure 13: Faces incorrectly identified as being the same by Humans. [Click here to view the full set of faces.](#)