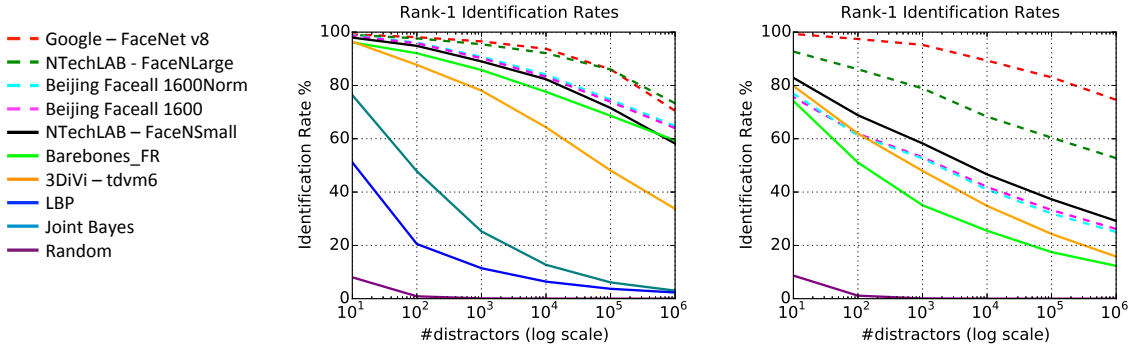
The MegaFace Benchmark: 1 Million Faces for Recognition at Scale

Ira Kemelmacher-Shlizerman Steven M. Seitz Daniel Miller Evan Brossard
Department of Computer Science and Engineering
University of Washington



(a) FaceScrub + MegaFace (b) FGNET + MegaFace

Figure 1. The MegaFace challenge evaluates identification and verification as a function of increasing number of gallery distractors (going from 10 to 1 Million). We use two different probe sets (a) FaceScrub–photos of celebrities, (b) FGNET–photos with a large variation in age per person. We present rank-1 identification of state of the art algorithms that participated in our challenge. On the left side of each plot is current major benchmark LFW scale (i.e., 10 distractors, see how all the top algorithms are clustered above 95%). On the right is mega-scale (with a million distractors). Observe that rates drop with increasing numbers of distractors, even though the probe set is fixed, and that algorithms trained on larger sets (dashed lines) generally perform better. Participate at: http://megaface.cs.washington.edu.

Abstract

Recent face recognition experiments on a major benchmark (LFW [15]) show stunning performance–a number of algorithms achieve near to perfect score, surpassing human recognition rates. In this paper, we advocate evaluations at the ***million scale*** *(LFW includes only 13K photos of 5K people). To this end, we have assembled the MegaFace dataset and created the first MegaFace challenge. Our dataset includes One Million photos that capture more than 690K different individuals. The challenge evaluates performance of algorithms with increasing numbers of "distractors" (going from 10 to 1M) in the gallery set. We present both identification and verification performance, evaluate performance with respect to pose and a persons age, and compare as a function of training data size (#photos and #people). We report results of state of the art and baseline algorithms. The MegaFace dataset, baseline code, and evaluation scripts, are all publicly released for further experimentations*[1].

[1]MegaFace data, code, and challenge can be found at: 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 [25, 29, 27] surpassing human performance on the leading Labeled Faces in the Wild (LFW) benchmark [15] 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 were searching for the suspect of the Boston marathon bombing [17]. 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. They also require training and testing on datasets that contain vast numbers of different people.

In this paper, we introduce the *MegaFace* dataset and benchmark to evaluate and encourage development of face recognition algorithms at scale. The goal of MegaFace 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 assembling the dataset are that 1) it should contain a million photos **"in the wild"**, i.e., with unconstrained pose, expression, lighting, and exposure, 2) be broad rather than deep, i.e., **contain many different people** rather than many photos of a small number of people, and most importantly 3) it will be **publicly available**, to enable benchmarking and distribution within the research community.

While recent face datasets have leveraged celebrity photos crawled from the web, such datasets have been limited to a few thousand unique individuals; it is challenging to find a million or more unique celebrities. Instead, we leverage Yahoo's recently released database of Flickr photos [31]. The Yahoo dataset includes 100M **creative commons** photographs and hence can be released for research. While these photos are unconstrained and do not target face recognition research per se, they capture a large number of faces. Our algorithm samples the Flickr set searching for faces while optimizing for large number of unique people via analysis of Flickr user IDs and group photos. MegaFace includes 1 Million photos of more than 690,000 unique subjects.

The MegaFace challenge evaluates how face recognition algorithms perform with a very large number of "distractors," i.e., individuals that are not in the probe set. MegaFace is used as the gallery; the two probe sets we use are FaceScrub [22] and FG-NET [7, 16]. We address fundamental questions and introduce the following key findings (Fig. 1):

- **How well do current face recognition algorithms scale?** Algorithms that achieve above 95% performance on LFW (equivalent of 10 distractors in our plots), achieve 35-75% identification rates with 1M distractors. Baselines (Joint Bayes and LBP) while achieving reasonable results on LFW drop to less than 10%.
- **Is the size of training data important?** We observe that algorithms that were trained on larger sets (top two are FaceNet that was trained on more than 500M photos of 10M people, and FaceN that was trained on 18M of 200K people) tend to perform better at scale. Interestingly, however, FaceN (trained on 18M) compares favorably to FaceNet (trained on 500M) on the FaceScrub set.
- **How does age affect recognition performance?** We found that the performance with 10 distractors for FGNET as a probe set is lower than for FaceScrub, and the drop off spread is much bigger (Fig. 1 (b)) . A deeper analysis also reveals that children (below age 20) are more challenging to recognize than adults, possibly due to training data availability, and that larger gaps in age (between gallery and probe) are similarly more challenging to recognize. These observations become evident by analyzing at large scale.
- **How does pose affect recognition performance?** Recognition drops for larger variation in pose between matching probe and gallery, and the effect is much more significant at scale.

In the following sections we describe how the MegaFace database was created, explain the challenge, and describe the outcomes.

2. Related Work

2.1. Benchmarks

Early work in face recognition focused on controlled datasets where subsets of lighting, pose, or facial expression were kept fixed, e.g., [10, 11]. 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 [15], Fig. 3, for a list of all the datasets), and thorough evaluations [13, 37]. A big challenge, however, was to collect photos of large number of individuals.

Large scale evaluations were previously performed on *controlled* datasets (visa photographs, mugshots, lab captured photos) by NIST [13], and report recognition results of 90% on 1.6 million people. However, these results are not representative of photos in the wild.

In 2007, Huang et al. [15] 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 [32] on Yahoo News photos. LFW captures celebrities photographed under unconstrained conditions (arbitrary lighting, pose, and expression) and it has been an amazing resource for the face analysis community (more than 1K citations). 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, e.g., [18, 3, 36, 34, 6, 22]. However, LFW remains the leading benchmark on which all state of the art recognition methods are evaluated and compared. Indeed, just in the last year a number of methods (11 methods at the time of writing this paper), e.g., [25, 28, 27, 29, 30] reported recognition rates above 99%+ [14] (better than human recognition rates estimated on the same dataset by [19]). The perfect recognition rate on LFW is 99.9% (it is not 100% since there are 5 pairs of photos that are mislabeled), and current top performer reports 99.77%. Recently, IJB-A dataset was released, it includes a large variation within an individual's photos, however is not large scale (26K photos total).

Stats \ Dataset	MegaFace (this paper)	CASIA-WebFace	LFW	PIPA	FaceScrub	YouTube Faces	Parkhi et al.	CelebFaces	DeepFace (Facebook)	NTechLab	FaceNet (Google)	WebFaces Wang et al.	IJB-A IAPRA
#photos	1,027,060	494,414	13K	60K	100K	3425 videos	2.6M	202K	4.4M	18.4M	>500M	80M	25,813
#subjects	**690,572**	10,575	5K	2K	500	1595	2.6K	10K	4K	200K	>10M	N/A	500
Source of photos	**Flickr**	Celebrity search	Yahoo News	**Flickr**	Celebrity search	Celebrities on YouTube	Celebrity search	Celebrity search	Internal	Internal	Internal	Web crawling	Internal
Public/private dataset	**Public**	**Public**	**Public**	**Public**	**Public**	**Public**	Private	Private	Private	Private	Private	Private	**Public**

Figure 2. Representative sample of recent face recognition datasets (in addition to LFW). Current public datasets include up to 10K unique people, and a total of 500K photos. Several companies have access to orders of magnitude more photos and subjects, these however are subject to privacy constraints and are not public. MegaFace (this paper) includes 1M photos of more than 690K unique subjects, collected from Flickr (from creative commons photos), and is available publicly.

2.2. Datasets

While, some companies have access to massive photo collections, e.g., Google in [25] trained on 200 Million photos of 8 Million people (and more recently on 500M of 10M), these datasets are not available to the public and were used only for training and not testing.

The largest public data set is CASIA-WebFace [36] that includes 500K photos of 10K celebrities, crawled from the web. While CASIA is a great resource, it contains only 10K individuals, and does not have an associated benchmark (i.e., it's used for training not testing).

Ortiz et al. [23] experimented with large scale identification from Facebook photos assuming there is more than one gallery photo per person. Similarly Stone et al. [26] show that social network's context improves large scale face recognition. Parkhi et al. [24] assembled a dataset of 2.6 Million of 2600 people, and used it for training (testing was done on the smaller scale LFW and YouTube Faces [34]). Wang et al. [33] propose a hierarchical approach on top of commercial recognizer to enable fast search in a dataset of 80 million faces. Finally, [4] experimented with a million distractors. Unfortunately, however, none of these efforts have produced publicly available datasets or public benchmarks.

2.3. Related Studies

Age-invariant recognition is an important problem that has been studied in the literature, e.g., [5, 20]. FG-NET [7] includes 975 photos of 82 people, each with several photos spanning many ages. More recently, Chen et al. [5] created a dataset of 160k photos of 2k celebrities across many ages, and Eidinger et al. [9] created a dataset of 27K photos of 2.3K Flickr user to facilitate age and gender recognition. Since most modern face recognition algorithms have not been evaluated for age-invariance we rectify this by including an FG-NET test (augmented with a million distractors) in our benchmark. In the future, datasets like [5, 9] can be easily incorporated into our benchmark.

Other recent studies have considered both identification as well as verification results on LFW [2, 30, 28, 27]. Finally, Best-Rowden et al. [2] performed an interesting Mechanical Turk study to evaluate human recognition rates on LFW and YouTube Faces datasets. They report that humans are better than computers when recognizing from videos due to additional cues, e.g., temporal information, familiarity with the subject (celebrity), workers' country of origin (USA vs. others), and also discovered errors in labeling of YouTube Faces via crowdsourcing. In the future, we will use this study's useful conclusions to help annotate MegaFace and create a training set in addition to the currently provided distractor set.

3. Assembling MegaFace

In this section, we provide an overview of the MegaFace dataset, how it was assembled, and its statistics. We created MegaFace to evaluate and drive the development of face recognition algorithms that work at scale. As motivated in Section 1, we sought to create a public dataset, free of licensing restrictions, that captures photos taken with unconstrained imaging conditions, and with close to a million unique identities. After exploring a number of avenues for data collection, we decided to leverage Yahoo's 100M Flickr set [31]. Yahoo's set was not created with face analysis in mind, however, it includes a very large number of faces and satisfies our requirements.

Optimizing for large number of unique identities. Our strategy for maximizing the number of unique identities is based on two techniques: 1) drawing photos from many different Flickr users—there are 500K unique user IDs—and 2) assuming that two or more faces appear in the same photo, they are likely different identities. Note that these assumptions do not need to be infallible, as our goal is to produce a very diverse distractor set–it is not a problem if we have a small number of photos of the same person. Our algorithm for detecting and downloading faces is as follows. We generated a list of images and user IDs in a round-robin fashion, by going through each of the 500K users and selecting the first photo with a face larger than 50×50 and adding it to the dataset. If the photo contains multiple faces

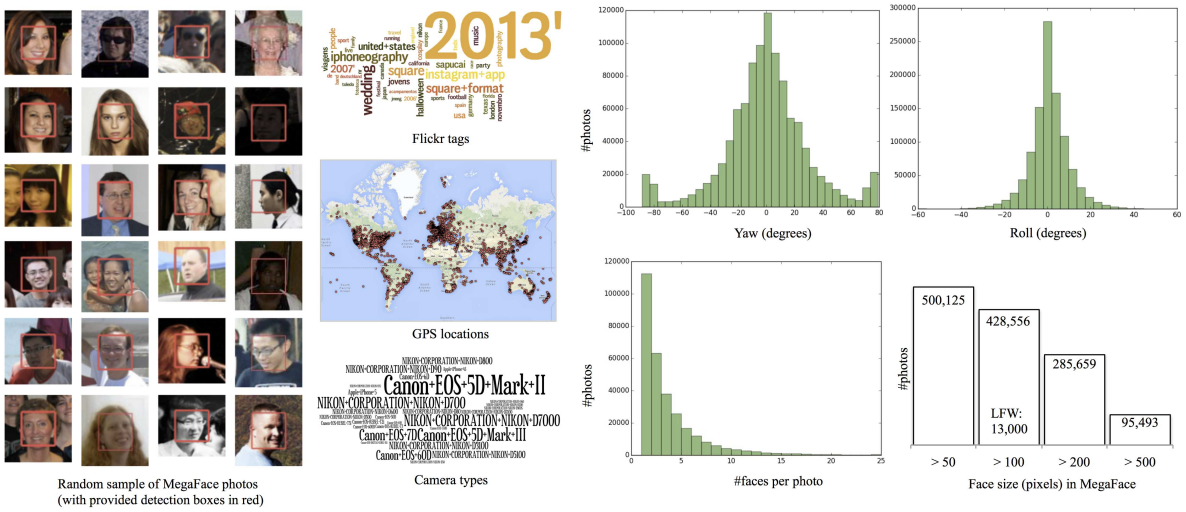


Figure 3. MegaFace statistics. We present randomly selected photographs (with provided detections in red), along with distributions of Flickr tags, GPS locations, and camera types. We also show the pose distribution (yaw and roll), number of faces per photograph, and number of faces for different resolutions (compared to LFW in which faces are approximately 100x100).

above that resolution, we add them all, given that they are different people with high probability. We then repeated this process (choosing the second, then the third, etc. photo from each user), until a sufficient number of faces were assembled. Based on our experiments face detection can have up to 20% false positive rate. Therefore, to ensure that our final set includes a million faces, the process was terminated once $1,296,079$ faces were downloaded. Once face detection was done, we ran additional stricter detection, and removed blurry faces. We assembled a total of $690,572$ faces in this manner that have a high probability of being unique individuals. While not guaranteed, the remaining 310K in our dataset likely also contain additional unique identities. Figure 3 presents a histogram of number of faces per photo.

Face processing. We downloaded the highest resolution available per photo. The faces are detected using the HeadHunter[2] algorithm by Mathias et al. [21], which reported state of the art results in face detection, and is especially robust to a wide range of head poses including profiles. We crop detected faces such that the face spans 50% of the photo height, thus including the full head (Fig. 3). We further estimate 49 fiducial points and yaw and pitch angles, as computed by the IntraFace[3] landmark model [35].

Dataset statistics. Figure 3 presents MegaFace's statistics:

- Representative photographs and bounding boxes. Observe that the photographs contain people from different countries, gender, variety of poses, glasses/no glasses, and many more variations.
- Distribution of Flickr tags that accompanied the downloaded photos. Tags range from 'instagram' to 'wedding,' suggesting a range of photos from selfies to high quality portraits (prominence of '2013' likely due to timing of when the Flickr dataset was released).
- GPS locations demonstrate photos taken all over the world.
- Camera types dominated by DSLRs (over mobile phones), perhaps correlated with creative commons publishers, as well as our preference for higher resolution faces.
- 3D pose information: more than 197K of the faces have yaw angles *larger* than ± 40 degrees. Typically unconstrained face datasets include yaw angles of *less* than ± 30 degrees.
- Number of faces per photo, to indicate the number of group photos.
- Face resolution: more than 50% (514K) of the photos in MegaFace have resolution more than 40 pixels inter-ocular distance (40 IOD corresponds to 100x100 face size, the resolution in LFW).

We believe that this dataset is extremely useful for a variety of research areas in recognition and face modeling, and we plan to maintain and expand it in the future. In the next section, we describe the MegaFace challenge.

4. The MegaFace Challenge

In this section, we describe the challenge and evaluation protocols. Our goal is to test performance of face recognition algorithms with up to a million distractors, i.e., faces of unknown people. In each test, a *probe* image is compared

[2] http://markusmathias.bitbucket.org/2014_eccv_face_detection/

[3] http://www.humansensing.cs.cmu.edu/intraface/

against a *gallery* of up to a million faces drawn from the Megaface dataset.

Recognition scenarios The first scenario is identification: given a probe photo, and a gallery containing at least one photo of the same person, the algorithm rank-orders all photos in the gallery based on similarity to the probe. Specifically, the probe set includes N people; for each person we have M photos. We then test each of the M photos (denote by i) per person by adding it the gallery of distractors and use each of the other $M-1$ photos as a probe. Results are presented with Cumulative Match Characteristics (CMC) curves– the probability that a correct gallery image will be chosen for a random probe by rank $= K$.

The second scenario is verification, i.e., a pair of photos is given and the algorithm should output whether the person in the two photos is the same or not. To evaluate verification we computed all pairs between the probe dataset and the Megaface distractor dataset. Our verification experiment has in total 4 billion negative pairs. We report verification results with ROC curves; this explores the trade off between falsely accepting non-match pairs and falsely rejecting match pairs.

Until now, verification received most of the focus in face recognition research since it was tested by the LFW benchmark [15]. Recently, a number of groups, e.g., [2, 30, 28, 27] also performed identification experiments on LFW. The relation between the identification and verification protocols was studied by Grother and Phillips [12] and DeCann and Ross [8]. In our challenge, we evaluate both scenarios with an emphasis on very large number of distractors. For comparison, testing identification on LFW is equivalent to 10 distractors in our challenge.

Probe set. MegaFace is used to create a gallery with a large number of distractors. For the probe set (testing known identities), we use two sets:

1. The FaceScrub dataset [22], which includes 100K photos of 530 celebrities, is available online. FaceScrub 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 which reduces possible bias, e.g., due to backgrounds and hair style [19], that may occur in LFW. For efficiency, the evaluation was done on a subset of FaceScrub which includes 80 identities (40 females and 40 males) by randomly selecting from a set of people that had more than 50 images each (from which 50 random photos per person were used).

2. The FG-NET aging dataset [7, 16]: it includes 975 photos of 82 people. For some of the people the age range in photos is more than 40 years.

Evaluation and Baselines. Challenge participants were asked to calculate their features on MegaFace, full FaceScrub, and FGNET. We provided code that runs identification and verification on the FaceScrub set. After the results were submitted by all groups we re-ran the experiments with FaceScrub and 3 different random distractor sets per gallery size. We further ran the FGNET experiments on all methods[4] and each of the three random MegaFace subsets per gallery size. The metric for comparison is L_2 distance. Participants were asked not to train on FaceScrub or FGNET. As a baseline, we implemented two simple recognition algorithms: 1) comparison by LBP [1] features–it achieves 70% recognition rates on LFW, and uses no training, 2) a Joint Bayesian (JB) approach represents each face as the sum of two Gaussian variables $x = \mu + \epsilon$ where μ is identity and ϵ is 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. JB algorithm was trained on the CASIA-WebFace dataset [36].

5. Results

This section describes the results and analysis of the challenge. Our challenge was released on Sep 30, 2015. Groups were given three weeks to finish their evaluations. More than 100 groups registered to participate. We present results from 5 groups that uploaded all their features by the deadline. We keep maintaining the challenge and data–currently 20 more groups are working on their submissions.

Participating algorithms In addition to baseline algorithms LBP, and Joint Bayes, we present results of the following methods (some provided more than 1 model):

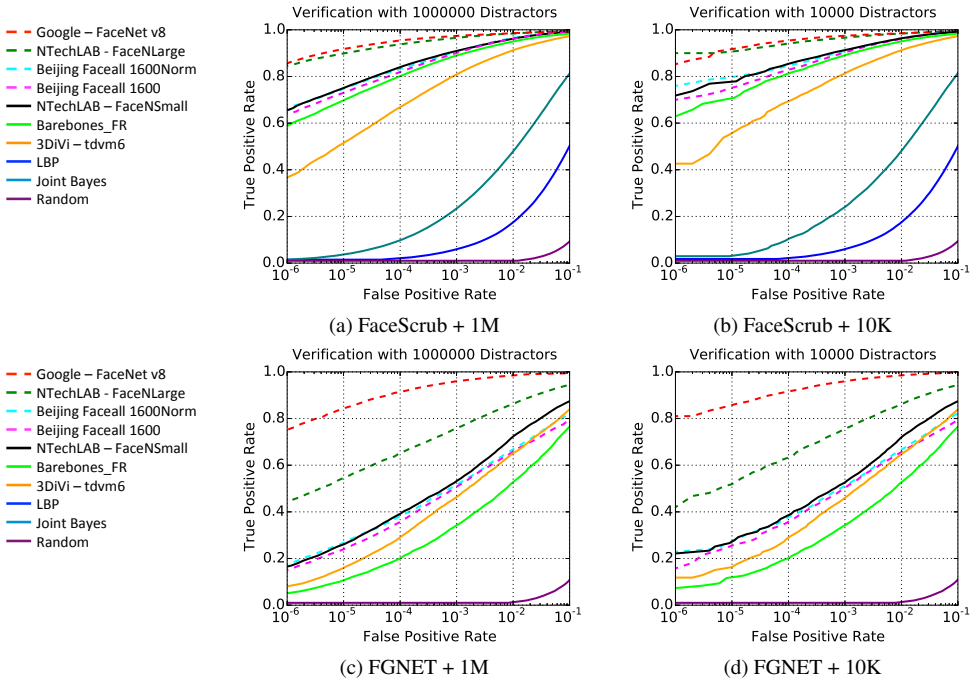
1. Google's FaceNet: achieves 99.6% on LFW, was trained on more than 500M photos of 10M people (newer version of [25]).
2. FaceAll (Beijing University of Post and Telecommunication), was trained on 838K photos of 17K people, and provided two types of features.
3. NTechLAB.com (FaceN algorithm): provided two models (small and large)–small was trained on 494K photos of 10K people, large on more than 18M of 200K.
4. BareBonesFR (University group): was trained on 365K photos of 5K people.
5. 3DiVi.com: was trained on 240K photos of 5K people.

Figure 4 summarizes the models, training sizes (240K-500M photos, 5K-10M people) and availability of the training data. Below we describe all the experiments and key conclusions.

[4] Google's FaceNet was ran by the authors since their features could not be uploaded due to licensing conditions

Group/ algorithm / Training data used	LBP	JointBayes	3DiVi	BareBonesFR (UMD)	FaceAll Beijing	Ntech Lab small model	Ntech Lab large model	FaceNet (Google)
#photos	0	494,414	240,000	365,495	838,776	494,414	18,435,445	>500M
#unique people	0	10,575	5,000	5,772	17,452	10,575	200K	>10M
Public/private dataset	N/A	Public (CASIA)	Private	Private	Private	Private	Private	Private

Figure 4. Number of training photos and unique people used by each participating method.



(a) FaceScrub + 1M (b) FaceScrub + 10K

(c) FGNET + 1M (d) FGNET + 10K

Figure 5. **Verification** performance with (a,c) 1 Million and (b,d) 10K distractors on both probe sets. Note the performance at low false accept rates (left side of each plot).

Algorithm / Probe set	Google FaceNet v8	NTechLAB FaceNLarge	Faceall Beijing 1600Norm	Faceall Beijing 1600	NTechLAB FaceNSmall	Barebones cnn	3DiVi tdvm6
FaceScrub	70.49	**73.30**	64.80	63.97	58.2	59.36	33.70
FGNET	**74.59**	52.71	25.02	26.15	29.16	12.31	15.77

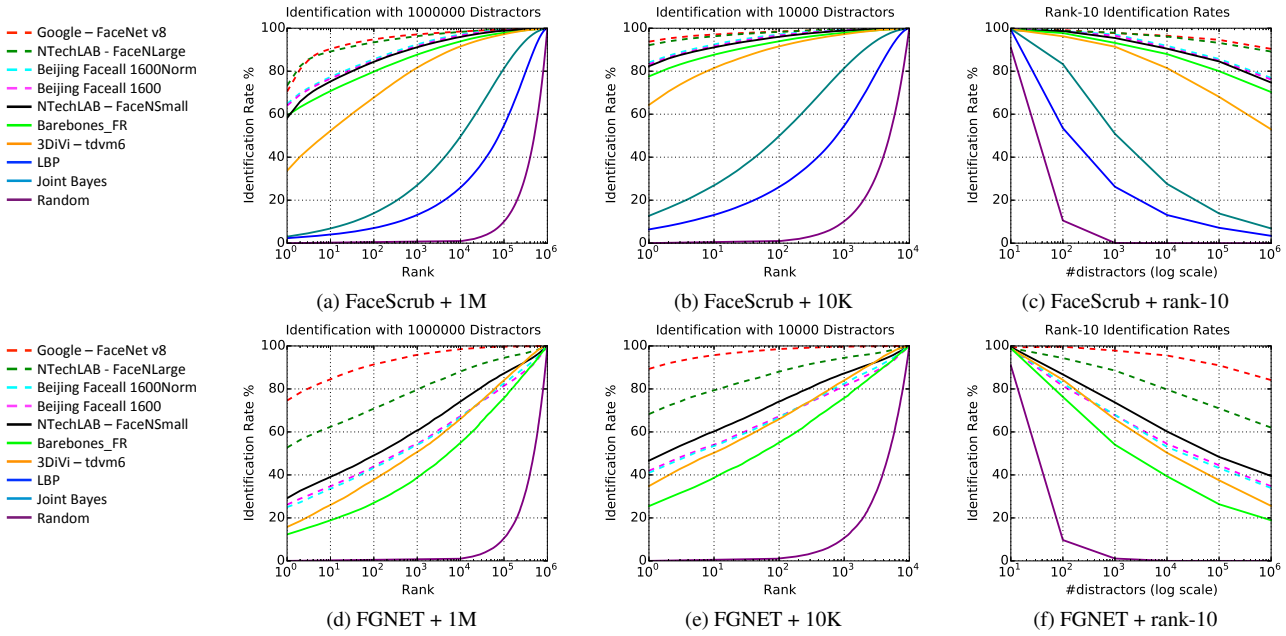
Figure 6. Rank-1 identification results (in%) with 1M distractors on the two probe sets.

Verification results. Fig. 5 shows results of the verification experiment for our two probe sets, (a) and (b) show results on FaceScrub and (c) and (d) on FGNET. We present results of one random fixed set of distractors per gallery size (see the other two in the supplementary).

We see that, for FaceScrub, at lower false accept rates the performance of algorithms drops by about 40% on average. FaceNet and FaceN lead with only about 15%. Interestingly, FaceN that was trained on 18M photos is able to achieve comparable results to FaceNet that was trained on 500M. 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. Results at LFW are typically reported at equal error rate which implies false accept rate of 1%-5% for top algorithms, while for a large set like MegaFace, only FAR of 10^-5 or 10^-6 is meaningful.

For FGNET the drop in performance is striking–about 60% for everyone but FaceNet, the latter achieving impressive performance across the board. One factor may be the type of training used by different groups (celebrities vs. photos across ages, etc.).

Verification rate stays similar when scaling up the gallery, e.g., compare (a) and (b). The intuition is that verification rate is normalized by the size of the dataset, so that if a probe face is matched incorrectly to 100 other faces in a 1000 faces dataset, 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 (FAR).



(a) FaceScrub + 1M (b) FaceScrub + 10K (c) FaceScrub + rank-10

(d) FGNET + 1M (e) FGNET + 10K (f) FGNET + rank-10

Figure 7. **Identification** performance for all methods with (a,d) 1M distractors and (b,e) 10K distractors, and (c,f) rank-10 for both probe sets. Fig. 1 also shows rank-1 performance as a function of number of distractors on both probe sets.

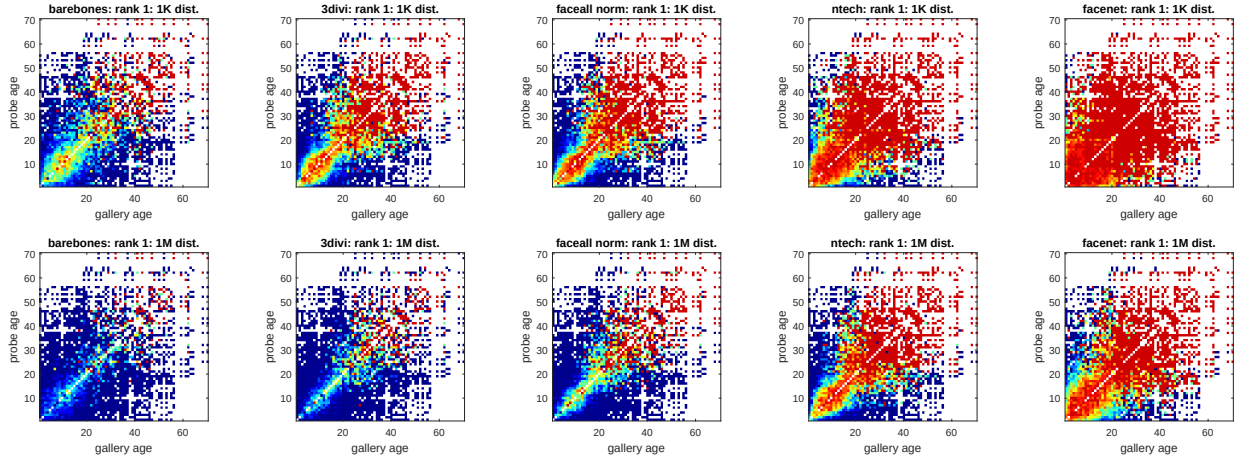


Figure 8. Analysis of rank-1 identification with respect to varying **ages** of gallery and probe. Columns represent five algorithms, rows 1K and 1M distractors. X-axis represents a person's age in the gallery photo and Y-axis age in the probe. The colors represent identification accuracy going from 0(=blue)–none of the true pairs were matched to 1(=red)–all possible combinations of probe and gallery were matched per probe and gallery ages. Lower scores on left and bottom indicate worse performance on children, and higher scores along the diagonal indicate that methods are better at matching across small age differences.

Identification results. In Fig. 7 we show the performance 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. (a,b,c) show performance for the FaceScrub dataset and (d,e,f) for FGNET. We observe that rates drop for all algorithms as the gallery size gets larger. This is visualized in Fig. 1, the actual accuracies are in Fig. 6. The curves also suggest that when evaluated on more than 1M distractors (e.g., 100M), rates will be even lower. Testing on FGNET **at scale** reveals a dramatic performance gap. All algorithms perform much worse, except for FaceNet that has a similar performance to its results on FaceScrub.

Training set size. Dashed lines in all plots represent algorithms that were trained on data larger than 500K photos and 20K people. We can see that these generally perform better than others.

Age. Evaluating performance using FGNET as a probe set also reveals a major drop in performance for most algorithms when attempting to match across differences in

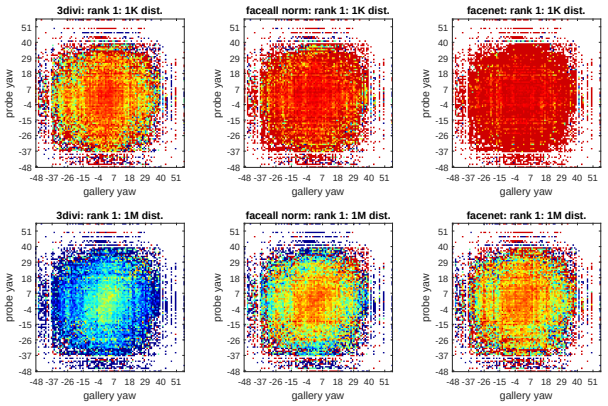


Figure 9. Analysis of rank-1 identification with varying **poses** of gallery and probe, for three algorithms. Top: 1K distractors, Bottom: 1M distractors. The colors represent identification accuracy going from 0 (blue) to 1 (red), where 0 means that none of the true pairs were matched, and 1 means that all possible combinations of probe and gallery were matched per probe and gallery ages. White color indicates combinations of poses that did not exist in our test set. We can see that evaluation at scale (bottom) reveals large differences in performance, which is not visible at smaller scale (top): frontal poses and smaller difference in poses is easier for identification.

(a) true positives

(b) false negatives

Figure 10. Examples pairs from FGNET using top performing FaceNet with 10 distractors. Each consecutive left-right pair of images is the same person. All algorithms match better with smaller age differences.

age. We present a number of results: Fig. 8 shows differences in performances with varying age across gallery and probe. Each column represents a different algorithm, rows present results for 1K and 1M distractors. Red colors indicate higher identification rate, blue lower rate. We make two key observations: 1) algorithms perform better when the difference in age between gallery and probe is small (along the diagonal), and 2) adults are more accurately matched than children, at scale. Fig. 10 shows examples of matched pairs (true positives and false negatives) using FaceNet and 10 distractors. Notice that false negatives have a bigger age gap relative to true positives. It is impressive, however, that the algorithm was able to these and many other true positives, given the variety in lighting, pose, and quality of the photo in addition to age changes.

Pose. Fig. 9 evaluates error in recognition as a function of difference in yaw between the probe and gallery. The results are normalized by the total number of pairs for each pose difference. We can see that recognition accuracy depends strongly on pose and this difference is revealed more prominently when evaluated at scale. Top row show results of three different algorithms (representative of others) with 1K distractors. Red colors indicate that identification is very high and mostly independent of pose. However, once evaluated at scale (bottom row) with 1M distractors we can see that variation across algorithms as well as poses is more dramatic. Specifically, similar poses identified better, and more frontal (center of the circle) poses are easier to recognize.

6. Discussion

An ultimate face recognition algorithm should perform with billions of people in a dataset. While testing with billions is still challenging, we have done the first step and created a benchmark of a million faces. MegaFace is available to researchers and we presented results from state of the art methods. Our key discoveries are 1) algorithms' performance degrades given a large gallery even though the probe set stays fixed, 2) testing at scale allows to uncover the differences across algorithms (which at smaller scale appear to perform similarly), 3) age differences across probe and gallery are still more challenging for recognition. We will keep maintaining and updating the MegaFace benchmark online, as well as, create more challenges in the future. Below are topics we think are exciting to explore. First, we plan to release all the detected faces from the 100M Flickr dataset. Second, companies like Google and Facebook have a head start due to availability of enormous amounts of data. We are interested to level the playing field and provide large **training** data to the research community that will be assembled from our Flickr data. Finally, the significant number of high resolution faces in our Flickr database will also allow to explore resolution in more depth. Currently, it is mostly untouched topic in face recognition literature due to lack of data.

Acknowledgments This work was funded in part by Samsung, Google's Faculty Research Award, and by NSF/Intel grant #1538613. We would like to thank the early challenge participants for great feedback on the baseline code and data.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006. 5

[2] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. 2014. 3, 5

[3] 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

[4] B. Bhattarai, G. Sharma, F. Jurie, and P. Pérez. Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In *Computer Vision-ECCV 2014 Workshops*, pages 160–172. Springer, 2014. 3

[5] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 3

[6] 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

[7] T. Cootes and A. Lanitis. The fg-net aging database, 2008. 2, 3, 5

[8] 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. 5

[9] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *Information Forensics and Security, IEEE Transactions on*, 9(12):2170–2179, 2014. 3

[10] A. Georghiades. Yale face database. *Center for computational Vision and Control at Yale University, http://cvc. yale. edu/projects/yalefaces/yalefa*, 1997. 2

[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 2

[12] 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. 5

[13] 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

[14] 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

[15] 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, 5

[16] I. Kemelmacher-Shlizerman, S. Suwajanakorn, and S. M. Seitz. Illumination-aware age progression. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3334–3341. IEEE, 2014. 2, 5

[17] 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

[18] 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

[19] 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, 5

[20] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *Information Forensics and Security, IEEE Transactions on*, 6(3):1028–1037, 2011. 3

[21] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014. 4

[22] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. *people*, 265(265):530. 2, 5

[23] E. G. Ortiz and B. C. Becker. Face recognition for web-scale datasets. *Computer Vision and Image Understanding*, 118:153–170, 2014. 3

[24] O. M. Parkhi, A. Vedaldi, A. Zisserman, A. Vedaldi, K. Lenc, M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, K. Lenc, et al. Deep face recognition. *Proceedings of the British Machine Vision*, 2015. 3

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*, 2015. 1, 2, 3, 5

[26] Z. Stone, T. Zickler, and T. Darrell. Autotagging facebook: Social network context improves photo annotation. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008. 3

[27] 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, 5

[28] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *arXiv preprint arXiv:1412.1265*, 2014. 2, 3, 5

[29] 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 IEEE Conference on*, pages 1701–1708. IEEE, 2014. 1, 2

[30] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. *arXiv preprint arXiv:1406.5266*, 2014. 2, 3, 5

[31] 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

[32] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 2

[33] D. Wang, C. Otto, and A. K. Jain. Face search at scale: 80 million gallery. *arXiv preprint arXiv:1507.07242*, 2015. 3

[34] 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, 3

[35] 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

[36] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2, 3, 5

[37] 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