MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, Jianfeng Gao

Microsoft Research
{yandong.guo,leizhang,yuxiao.hu,xiaohe,jfgao}@microsoft.com

Abstract. In this paper, we design a benchmark task and provide the associated datasets for recognizing face images and link them to corresponding entity keys in a knowledge base. More specifically, we propose a benchmark task to recognize one million celebrities from their face images, by using all the possibly collected face images of this individual on the web as training data. The rich information provided by the knowledge base helps to conduct disambiguation and improve the recognition accuracy, and contributes to various real-world applications, such as image captioning and news video analysis. Associated with this task, we design and provide concrete measurement set, evaluation protocol, as well as training data. We also present in details our experiment setup and report promising baseline results. Our benchmark task could lead to one of the largest classification problems in computer vision. To the best of our knowledge, our training dataset, which contains 10M images in version 1, is the largest publicly available one in the world.

Keywords: Face recognition, large scale, benchmark, training data, celebrity recognition, knowledge base

1 Introduction

In this paper, we design a benchmark task as to recognize one million celebrities from their face images and identify them by linking to the unique entity keys in a knowledge base. We also construct associated datasets to train and test for this benchmark task. Our paper is mainly to close the following two gaps in current face recognition, as reported in [1]. First, there has not been enough effort in determining the identity of a person from a face image with disambiguation, especially at the web scale. The current face identification task mainly focuses on finding similar images (in terms of certain types of distance metric) for the input image, rather than answering questions such as "who is in the image?" and "if it is Anne in the image, which Anne?". This lacks an important step of "recognizing". The second gap is about the scale. The publicly available datasets are much smaller than that being used privately in industry, such as Facebook [2, 3] and Google [4], as summarized in Table 1. Though the research in face

This paper is published at ECCV 2016.

recognition highly desires large datasets consisting of many distinct people, such large dataset is not easily or publicly accessible to most researchers. This greatly limits the contributions from research groups, especially in academia.

Our benchmark task has the following properties. First, we define our face recognition as to determine the identity of a person from his/her face images. More specifically, we introduce a **knowledge base** into face recognition, since the recent advance in knowledge bases has demonstrated incredible capability of providing accurate identifiers and rich properties for celebrities. Examples include Satori knowledge graph in Microsoft and "freebase" in [5]. Our face recognition task is demonstrated in Fig. 1.

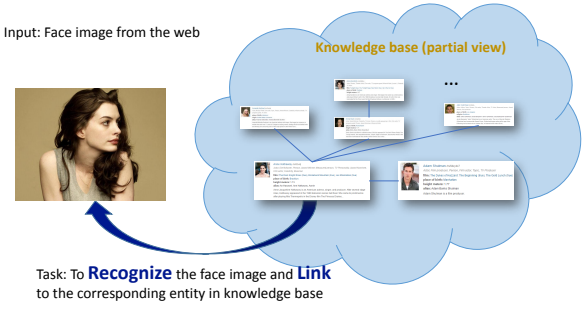


Fig. 1. An example of our face recognition task. Our task is to recognize the face in the image and then link this face with the corresponding entity key in the knowledge base. By recognizing the left image to be "Anne Hathaway" and linking to the entity key, we know she is an American actress born in 1982, who has played Mia Thermopolis in The Princess Diaries, not the other Anne Hathaway who was the wife of William Shakespeare. Input image is from the web. [2]

Linking the image with an entity key in the knowledge base, rather than an isolated string for a person's name naturally solves the disambiguation issue in the traditional face recognition task. Moreover, the linked entity key is associated with rich and comprehensive property information in the knowledge base, which makes our task more similar to human behavior compared with traditional face identification, since retrieving the individual's name as well as the associated information naturally takes place when humans are viewing a face image. The rich information makes our face recognition task practical and beneficial to many real applications, including image search, ranking, caption generation, image deep understanding, etc.

Second, our benchmark task targets at recognizing **celebrities**. Recognizing celebrities, rather than a pre-selected private group of people, represents public interest and could be directly applied to a wide range of real scenarios. More-

[1] Image resource: http://www.hdwallpapers.in/anne_hathaway_2-wallpapers.html, retrieved by image.bing.com.

over, only with popular celebrities, we can leverage the existing information (e.g. name, profession) in the knowledge base and the information on the web to build a large-scale dataset which is publicly available for training, measurement, and re-distributing under certain licenses. The security department may have many labeled face images for criminal identification, but the data can not be publicly shared.

Third, we select **one million** celebrities from freebase and provide their associated entity keys, and encourage researchers to build recognizers to identify each people entity. Considering each entity as one class may lead to, to the best of our knowledge, the largest classification problem in computer vision. The clear definition and mutually exclusiveness of these classes are supported by the unique entity keys and their associated properties provided by the knowledge base, since in our dataset, there are a significant amount of celebrities having same/similar names. This is different from generic image classification, where to obtain a large number of exclusive classes with clear definition itself is a challenging and open problem [6].

The large scale of our problem naturally introduces the following attractive challenges. With the increased number of classes, the inter-class variance tends to decrease. There are celebrities look very similar to each other (or even twins) in our one-million list. Moreover, large intra-class variance is introduced by popular celebrities with millions of images available, as well as celebrities with very large appearance variation (e.g., due to age, makeups, or even sex reassignment surgery).

In order to evaluate the performance of our benchmark task, we provide concrete measurement set and evaluation protocol. Our measurement set consists of images for a subset of celebrities in our one-million celebrity list. The celebrities are selected in a way that, our measurement set mainly focuses on popular celebrities to represent the interest of real application and users, while the measurement set still maintains enough (about 25%) tail celebrities to encourage the performance on celebrity coverage. We manually label images for these celebrities carefully. The correctness of our labeling is ensured by deep research on the web content, consensus verification, and multiple iterations of carefully review. In order to make our measurement more challenging, we blend a set of distractor images with this set of carefully labeled images. The distractor images are images of other celebrities or ordinary people on the web, which are mainly used to hide the celebrities we select in the measurement.

Along with this challenging yet attractive large scale benchmark task proposed, we also provide a very large training dataset to facilitate the task. The training dataset contains about 10M images for 100K top celebrities selected from our one-million celebrity list in terms of their web appearance frequency. Our training data is, to the best of our knowledge, the largest publicly available one in the world, as shown in Table 1. We plan to further extend the size in the near future. For each of the image in our training data, we provide the thumbnail of the original image and cropped face region from the original

image (with/without alignment). This is to maximize the convenience for the researchers to investigate using this data.

With this training data, we trained a convolutional deep neural network with the classification setup (by considering each entity as one class). The experimental results show that without extra effort in fine-tuning the model structure, we recognize 44.2% of the images in the measurement set with the precision 95% (hard case, details provided in section 4). We provide the details of our experiment setup and experimental results to serve as a very promising baseline in section 4.

Contribution Summary Our contribution in this paper is summarized as follows.

- We design a benchmark task: to recognize one million celebrities from their face images, and link to their corresponding entity keys in freebase [5].
- We provide the following datasets,[2]
  - One million celebrities selected from freebase with corresponding entity keys , and a snapshot for freebase data dumps;
  - Manually labeled measurement set with carefully designed evaluation protocol;
  - A large scale training dataset, with face region cropped and aligned (to the best of our knowledge, the largest publicly available one).
- We provide promising baseline performance with our training data to inspire more research effort on this task.

Our benchmark task could lead to a very large scale classification problem in computer vision with meaningful real applications. This benefits people in experimenting different recognition models (especially fine-grained neural network) with the given training/testing data. Moreover, we encourage people to bring in more outside data and evaluate experimental results in a separate track.

2 Related works

Typically, there are two types of tasks for face recognition. One is very well-studied, called face verification, which is to determine whether two given face images belong to the same person. Face verification has been heavily investigated. One of the most widely used measurement sets for verification is Labeled Faces in the Wild (LFW) in [7, 8], which provides 3000 matched face image pairs and 3000 mismatched face image pairs, and allows researchers to report verification accuracy with different settings. The best performance on LFW datasets has been frequently updated in the past several years. Especially, with the "unrestricted, labeled outside data" setting, multiple research groups have claimed higher accuracy than human performance for verification task on LFW [4, 9].

[2] Instructions and download links: http://msceleb.org

Recently, the interest in the other type of face recognition task, face identification, has greatly increased [9–11, 3]. For typical face identification problems, two sets of face images are given, called gallery set and query set. Then the task is, for a given face image in the query set, to find the most similar faces in the gallery image set. When the gallery image set only has a very limited number (say, less than five) of face images for each individual, the most effective solution is still to learn a generic feature which can tell whether or not two face images are the same person, which is essentially still the problem of face verification. Currently, the MegaFace in [11] might be one of the most difficult face identification benchmarks. The difficulty of MegaFace mainly comes from the up-to one million distractors blended in the gallery image set. Note that the query set in MegaFace are selected from images from FaceScrub [12] and FG-NET [13], which contains 530 and 82 persons respectively.

Several datasets have been published to facilitate the training for the face verification and identification tasks. Examples include LFW [7, 8], Youtube Face Database (YFD) [14], CelebFaces+ [15], and CASIA-WebFace [16]. In LFW, 13000 images of faces were collected from the web, and then carefully labeled with celebrities' names. The YFD contains 3425 videos of 1595 different people. The CelebFace+ dataset contains 202,599 face images of 10,177 celebrities. People in CelebFaces+ and LFW are claimed to be mutually exclusive. The CASIA-WebFace [16] is currently the largest dataset which is publicly available, with about 10K celebrities, and 500K images. A quick summary is listed in Table 1.

Table 1. Face recognition datasets

Dataset	Available	people	images
IJB-A [17]	public	500	5712
LFW [7, 8]	public	5K	13K
YFD [14]	public	1595	3425 videos
CelebFaces [15]	public	10K	202K
CASIA-WebFace [16]	public	10K	500K
Ours	**public**	100**K**	about 10**M**
Facebook	private	4K	4400K
Google	private	8M	100-200M

As shown in Table 1, our training dataset is considerably larger than the publicly available datasets. Another uniqueness of our training dataset is that our dataset focuses on facilitating our celebrity recognition task, so our dataset needs to cover as many popular celebrities as possible, and have to solve the data disambiguation problem to collect right images for each celebrity. On the other hand, the existing datasets are mainly used to train a generalizable face feature, and celebrity coverage is not a major concern for these datasets. Therefore, for the typical existing dataset, if a name string corresponds to multiple

celebrities (e.g., Mike Smith) and would lead to ambiguous image search result, these celebrities are usually removed from the datasets to help the precision of the collected training data [18].

3 Benchmark construction

Our benchmark task is to recognize one million celebrities from their face images, and link to their corresponding entity keys in the knowledge base. Here we describe how we construct this task in details.

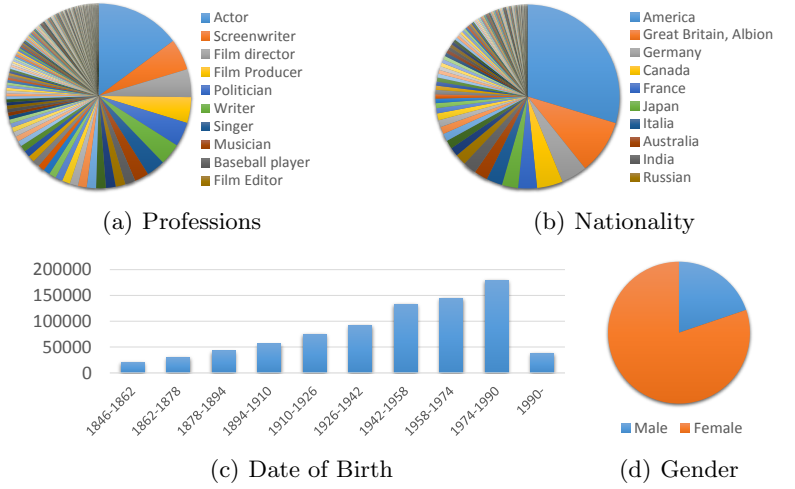
3.1 One million celebrity list

We select one million celebrities to recognize from a knowledge graph called freebase [5], where each entity is identified by a unique key (called machine identifier, MID in freebase) and associated with rich properties. We require that the entities we select are human beings in the real world and have/had public attentions.

The first step is to select a subset of entities (from freebase [5]) which correspond to real people using the criteria in [1]. In freebase, there are more than 50 million topics capsulated in about 2 billion triplets. Note that we don't include any person if his/her facial appearance is unknown or not clearly defined.

The second step is to rank all the entities in the above subset according to the frequency of their occurrence on the web [1]. We select the top one million entities to form our celebrity list and provide their entity keys (MID) in freebase. We concern the public attention (popularity on the web) for two reasons. First, we want to align our benchmark task with the interest of real applications. For applications like image search, image annotations and deep understanding, and image caption generation, the recognition of popular celebrities would be more attractive to most of the users than ordinary people. Second, we include popular celebrities so that we have better chance to obtain multiple authority images for each of them to enable our training, testing, and re-distributing under certain licenses.

We present the distribution of the one million celebrities in different aspects including profession, nationality, age, and gender. In our one million celebrity list, we include persons with more than 2000 different professions (Fig. 2 (a)), and come from more than 200 distinct countries/regions (Fig. 2 (b)), which introduces a great diversity to our data. We cover all the major races in the world (Caucasian, Mongoloid, and Negroid). Moreover, as shown in Fig. 2 (c), we cover a large range of ages in our list. Though we do not manually select celebrities to make the profession (or gender, nationality, age) distribution uniform, the diversity (gender, age, profession, race, nationality) of our celebrity list is guaranteed by the large scale of our dataset. This is different from [17], in which there are about 500 subjects so the manual balancing over gender distribution is inevitable.



(a) Professions (b) Nationality

(c) Date of Birth (d) Gender

Fig. 2. Distribution of the properties of the celebrities in our one-million list in different aspects. The large scale of our dataset naturally introduces great diversity. As shown in (a) and (b), we include persons with more than 2000 different professions, and come from more than 200 distinct countries/regions. The figure (c) demonstrates that we don't include celebrities who were born before 1846 (long time before the first roll-film specialized camera "Kodak" was invented [19]) and covers celebrities of a large variance of age. In (d), we notice that we have more females than males in our one-million celebrity list. This might be correlated with the profession distribution in our list.

Note that our property statistics are limited to the availability of freebase information. Some celebrities in our one million list do not have complete properties. If a certain celebrity does not have property A available in freebase, we do not include this celebrity for the statistic calculation of the property A.

3.2 Celebrity selection for measurement

In order to evaluate the recognition performance on the one million celebrities obtained in the last subsection, we build up a measurement set which includes a set of carefully labeled images blended with another set of randomly selected face images as distractors. The measurement set construction is described in details in the following subsections, while the evaluation protocol is described in Section 4.

For the labeled images, we sample a subset of celebrities [3] from the one-million celebrity list due to limited labeling resource. The sampling weight is designed in a way that, our measurement set mainly focuses on top celebrities (rank among the top in the occurrence frequency list) to represent the interest

[3] Currently there are 1500. We will increase the number of celebrities in our measurement set in the future.

of real applications and users, yet maintain a certain amount of tail celebrities (celebrities not mentioned frequently on the web, e.g., from 1 to 10 times in total) to guarantee the measurement coverage over the one-million list.

More specifically, let f_i denote the number of documents mentioned the i^{th} celebrity on the web. Following the method in [1], we set the probability for the i^{th} celebrity to get selected to be proportional to f'_i, defined as,

$$f'_i = f_i^{\frac{1}{\sqrt{5}}}, \tag{1}$$

where the exponent $1/\sqrt{5}$ is obtained empirically to include more celebrities with small f.

Though it seems to be a natural solution, we do not set the sampling weights to be proportional to f_i, since this option will make our measurement set barely contain any celebrities from the bottom 90% in our one-million list (ordered by f_i). The reason is that the distribution of f is very long-tailed. More than 90% of the celebrities have f smaller than 30, while the top celebrities have f larger than one million. We need to include sufficient number of tail celebrities to encourage researchers to work on the hard cases to improve the performance from the perspective of recognition coverage. This is the reason that we applied the adjustment in (1).

With the sampling weight f' in (1) applied, our measurement set still mainly focuses on the most popular celebrities, while about 25% of the celebrities in our measurement set come from the bottom 90% in our one-million celebrity list (ordered by f). If we do not apply the adjustment in (1), but just use f as the sampling weight, less than 10% of the celebrities in the measurement set come from the bottom 90% in our one-million celebrity list.

Since the list of the celebrities in our measurement set is not exposed [4], and our measurement set contains 25% of the celebrities in our measurement set come from the bottom 90%, researchers need to include as many celebrities as possible (not only the popular ones) from our one-million list to improve the performance of coverage. This pushes the scale of our task to be very large.

3.3 Labeling for measurement

After we have the set of celebrities for measurement, we provide two images for each of the celebrity. The correctness of our image labeling is ensured by deep research on the web content, multiple iterations of carefully review, and very rigorous consensus verification. Details are listed as follows.

Scraping Scraping provides image candidates for each of the celebrities selected for the measurement set. Though in the end we provide only two images per celebrity for evaluation, we scraped about 30 images per celebrities. During the scraping procedure, we applied different search queries, including the celebrity's

[4] We publish the images for 500 celebrities, called development set, while hold the rest 1000 for grand challenges.

name, name plus profession, and names in other languages (if available). The advantages of introducing multiple variations of the query used for each celebrity is that with multiple queries, we have better chance to capture the images which are truly about the given celebrity. Moreover, the variation of the query and scraping multiple images also brings in the diversity to the images for the given celebrity. Especially for the famous celebrities, the top one image returned by search engine is typically his/her representative image (frontal facial image with high quality), which is relatively easier to recognize, compared with the other images returned by the search engine. We increase the scraping depth so that we have more diverse images to be recognized for each of the celebrity.

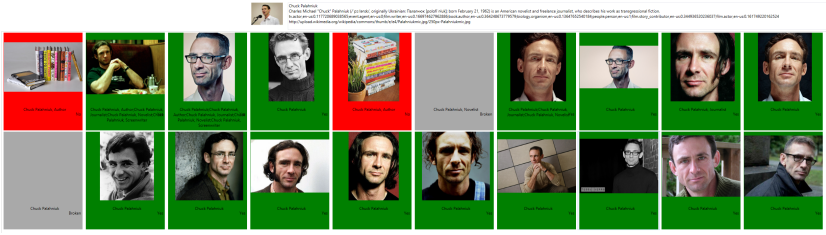


Fig. 3. Labeling GUI for "Chuck Palhniuk". (partial view) As shown in the figure, in the upper right corner, a representative image and a short description is provided. For a given image candidate, judge can label as "not for this celebrity" (red), "yes for this celebrity" (green), or "broken image" (dark gray).

Label Labeling picks up the images which are truly about the given celebrity. As shown in Fig.3, for each given celebrity, we (all the authors) manually label **all** the scraped image candidates to be truly about this celebrity or not. Extreme cautious was applied. We have access to the page which contains the scraped image to be labeled. Whenever needed, the judge (the authors) is asked to visit the original page with the scraped image and read the page content to guide his/her labeling. The rich information on the original page benefits the quality of the labeling, especially for a lot of the hard cases. Each of the image-celebrity entity pair was judged by at least two persons. Whenever there is a conflict, the two judges review together and provide the final decision based on verbal discussion. In total, we have about $30K$ images labeled, spent hundreds of hours.

In our measurement set, we select two images for each of the celebrity to keep the evaluation cost low. We have two subset (each of them have the same celebrity list), described as follows.

- **Random set**
  The image in this subset is randomly selected from the labeled images. One image per celebrity. This set reveals how many celebrities are truly covered by the models to be tested.
- **Hard set**

The image in this subset is the one (from the labeled images) which is the most different from any images in the training dataset. One image per celebrity. This set is to evaluate the generalization ability of the model.

Then, we blend the labeled images with images from other celebrities or ordinary people. The evaluation protocol is introduced in details in the next section.

4 Celebrity recognition

In this section, we set up the evaluation protocol for our benchmark task. Moreover, in order to facilitate the researchers to work on this problem, we provide a training dataset which is encouraged (optional) to use. We also present the baseline performance obtained by using our provided training data. We also encourage researchers to train with outside data and evaluate in a separate track.

4.1 Evaluation Protocol

We evaluate the performance of our proposed recognition task in terms of precision and coverage (defined in the following subsection) using the settings described as follows.

Setup We setup our evaluation protocol as follows. For a model to be tested, we collect the model prediction for both the labeled image and distractors in the measurement set. Note that we don't expose which images in the measurement are labeled ones or which are distractors. This setup avoids human labeling to the measurement set, and encourages researchers to build a recognizer which could robustly distinguish one million (as many as possible) people faces, rather than focusing merely on a small group of people.

Moreover, during the training procedure, if the researcher leverages outside data for training, we do not require participants to exclude celebrities in our measurement from the training data set. Our measurement still evaluate the generalization ability of our recognition model, due to the following reasons. There are one million celebrities to be recognized in our task, and there are millions of images for some popular celebrities on the web. It is practically impossible to include all the images for every celebrity in the list. On the other hand, according to section 4.2, the images in our measurement set is typically not the representative images for the given celebrity (e.g., the top one searching result). Therefore the chance to include the measurement images in the training set is relatively low, as long as the celebrity list in the measurement set is hidden. This is different from most of the existing face recognition benchmark tasks, in which the measurement set is published and targeted on a small group of people. For these traditional benchmark tasks, the evaluation generalization ability relies on manually excluding the images (from the training set) of all the persons in the measurement set (This is mainly based on the integrity of the participants).

Evaluation metric In the measurement set, we have n images, denoted by $\{x_i\}_{i=1}^n$. The first m images $\{x_i|i = 1, 2, 3, ..., m\}$ are the labeled images for our selected celebrities, while the rest $\{x_i|i = m+1, ..., n\}$ are distractors. Note that we hide the order of the images in the measurement set.

For the i^{th} image, let $g(x_i)$ denote the ground truth label (entity key obtained by labeling). For any model to be tested, we assume the model to output $\{\hat{g}(x_i), c(x_i)\}$ as the predicted entity key of the i^{th} image, and its corresponding prediction confidence. We allow the model to perform rejection. That is, if $c(x_i) < t$, where t is a preset threshold, the recognition result for image x_i will be ignored. We define the precision with the threshold t as,

$$P(t) = \frac{|\{x_i|\hat{g}(x_i) = g(x_i) \wedge c(x_i) \geq t, i = 1, 2, ..., m\}|}{|\{x_i|c(x_i) \geq t, i = 1, 2, ..., m\}|}, \tag{2}$$

where the nominator is the number of the images of which the prediction is correct (and confidence score is larger than the threshold). The denominator is the number of images (within the set $\{x_i\}_{i=1}^m$) which the model does have prediction (not reject to recognize).

The coverage in our protocol is defined as

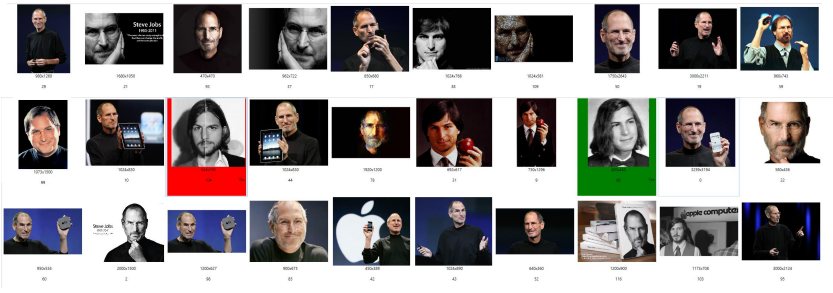
$$C(t) = \frac{|\{x_i|c(x_i) \geq t, i = 1, 2, ..., m\}|}{m} \tag{3}$$

For each given t, a pair of precision $P(t)$ and coverage $C(t)$ can be obtained for the model to be tested. The precision $P(t)$ is a function of $C(t)$. Our major evaluation metric is the maximum of the coverage satisfying the condition of precision, $P(t) \geq P_{min}$. The value of P_{min} is 0.95 in our current setup. Other metrics and analysis/discussions are also welcomed to report. The reason that we prefer a fixed high precision and measure the corresponding coverage is because in many real applications high precision is usually more desirable and of greater value.

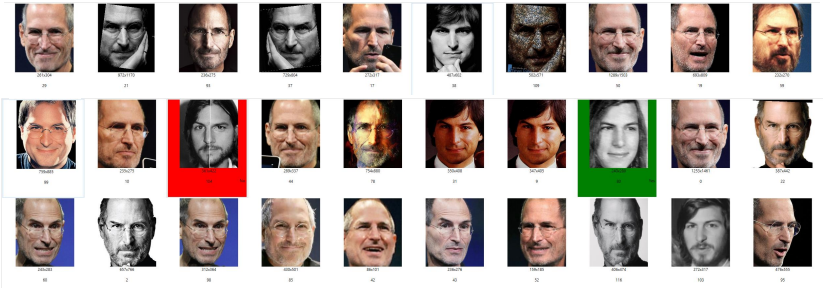
4.2 Training dataset

In order to facilitate the above face recognition task we provide a large training dataset. This training dataset is prepared by the following two steps. First, we select the top 100K entities from our one-million celebrity list in terms of their web appearance frequency. Then, we retrieve approximately 100 images per celebrity from popular search engines.

We do not provide training images for the entire one-million celebrity list for the following considerations. First, limited by time and resource, we can only manage to prepare a dataset of top 100K celebrities as a v1 dataset to facilitate the participants to quickly get started. We will continuously extend the dataset to cover more celebrities in the future. Moreover, as shown in the experimental results in the next subsection, this dataset is already very promising to use. Our training dataset covers about 75% of celebrities in our measurement set, which implies that the upper bound of recognition recall rate based on the provided



(a) Original Image



(b) Aligned Face Image

Fig. 4. Examples (subset) of the training images for the celebrity with entity key m.06y3r (Steve Jobs). The image marked with a green rectangle is claimed to be Steve Jobs when he was in high school. The image marked with a red rectangle is considered as a noise sample in our dataset, since it is synthesized by combining one image of Steve Jobs and one image of Ashton Kutcher, who is the actor in the movie "Jobs".

training data cannot exceed 75%. Therefore, we also encourage the participants, especially who are passionate to break this 75% upper bound to treat the dataset development as one of the key problems in this challenge, and bring in outside data to get higher recognition recall rate and compare experimental results in a separate track. Especially, we encourage people label their data with entity keys in the freebase snapshot we provided and publish, so that different dataset could be easily united to facilitate collaboration.

On example in our training dataset is shown in Figure 4. As shown in the figures, same celebrity may look very differently in different images. In Figure 4, we see images for Steve Jobs (m.06y3r) when he was about 20/30 years old, as well as images when he was about 50 years old. The image at row 2, column 8 (in green rectangle) in Figure 4 is claimed to be Steve Jobs when he was in high school. Notice that the image at row 2, column 3 in Figure 4, marked with red rectangle is considered as a noise sample in our dataset, since this image was synthesized by combining one image of Steve Jobs and one image of Ashton Kutcher, who is the actor in the movie "Jobs".

As we have mentioned, we do not manually remove the noise in this training data set. This is partially because to prepare training data of this size is beyond the scale of manually labeling. In addition, we have observed that the state-of-the-art deep neural network learning algorithm can tolerate a certain level of noise in the training data. Though for a small percentage of celebrities their image search result is far from perfect, more data especially more individuals covered by the training data could still be of great value to the face recognition research, which is also reported in [18]. Moreover, we believe that data cleaning, noisy label removal, and learning with noisy data are all good and real problems that are worth of dedicated research efforts. Therefore, we leave this problem open and do not limit the use of outside training data.

4.3 Baseline

There are typically two categories of methods to recognize people from face images. One is template-based. For methods in this category, a gallery set which contains multiple images for the targeted group of people is pre-built. Then, for the given image in the query set, the most similar image(s) in the gallery set (according to some certain metrics or in pre-learned feature space) is retrieved, and the annotation of this/these similar images are used to estimate the identity of the given query image. When the gallery is not very large, this category of methods is very convenient for adding/removing entities in the gallery since the face feature representation could be learned in advance. However, when the gallery is large, a complicated index needs to be built to shorten the retrieval time. In this case, the flexibility of adding/removing entities for the methods in this category vanishes. Moreover, the accuracy of the template-based methods highly relies on the annotation accuracy in the gallery set. When there are many people in the targeted group, accurate annotation is beyond human effort and could be a very challenging problem itself.

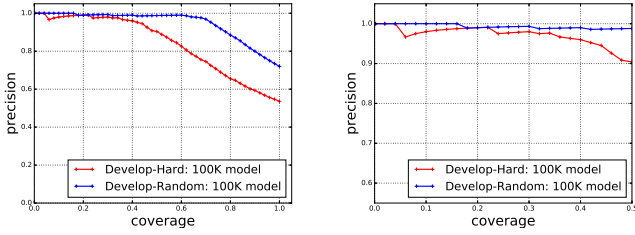
We choose the second category, which is a model-based method. More specifically, we model our problem as a classification problem and consider each celebrity as a class.

In our experiment, we trained a deep neural network following the network structure in [20]. Training a deep neural network for 100K celebrities is not a trivial task. If we directly train the model from scratch, it is hard to see the model starts to converge even after a long run due to the large number of categories. To address this problem, we started from training a small model for 500 celebrities, which have the largest numbers of images for each celebrity. In addition, we used the pre-trained model from [20] to initialize this small model. This step is optional, but we observed that it helps the training process converge faster. After $50,000$ iterations, we stopped to train this model, and used it as a pre-trained model to initialize the full model of 100K celebrities. After $250,000$ iterations, with learning rate decreased from the initial value 0.01 to 0.001 and 0.0001 after $100,000$ and $200,000$ iterations, the training loss decrease becomes very slow and indiscernible. Then we stopped the training and used the last model snapshot to evaluate the performance of celebrity recognition on our measurement set. The

experimental results (on the published 500 celebrities) are shown in Fig. 5 and Table 2.

Table 2. Experimental results on the 500 published celebrities

	Coverage@Precision 99%	Coverage@Precision 95%
Hard Set	0.052	0.442
Random Set	0.606	0.728



(a) Original Curve (b) Zoom-in

Fig. 5. Precision-coverage curve with our baseline model

The promising results can be attributed to the deep neural network capability and the high quality of image search results thanks for years of improvement in image search engines. However, the curves also shows that the task is indeed very challenge. To achieve both high precision and high recall, a great amount of research efforts need to be spent on data collection, cleaning, learning algorithm, and model generalization, which are valuable problems to computer vision researchers.

5 Discussion and Future work

In this paper, we have defined a benchmark task which is to recognize one million celebrities in the world from their face images, and link the face to a corresponding entity key in a knowledge base. Our face recognition has the property of disambiguation, and close to the human behavior in recognizing images. We also provide concrete measurement set for people to evaluate the model performance easily, and provide, to the best of our knowledge, the largest training dataset to facilitate research in the area.

Beyond face recognition, our datasets could inspire other research topics. For example, people could adopt one of the cutting-edge unsupervised/semi-

supervised clustering algorithms [21] [22] [23] [24] on our training dataset, and/or develop new algorithms which can accurately locate and remove outliers in a large, real dataset. Another interesting topic is the to build estimators to predict a person's properties from his/her face images. For example, the images in our training dataset are associated with entity keys in knowledge base, of which the gender information (or other properties) could be easily retrieved. People could train a robust gender classifier for the face images in the wild based on this large scale training data. We look forward to exciting research inspired by our training dataset and benchmark task.

References

1. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: Challenge of recognizing one million celebrities in the real world. In: IS&T International Symposium on Electronic Imaging. (2016)
2. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR). (June 2014)
3. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Web-scale training for face identification. In: Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR), IEEE (2015) 2746–2754
4. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR). (June 2015)
5. Google: Freebase data dumps. https://developers.google.com/freebase/data (2015)
6. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3) (2015) 211–252
7. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (October 2007)
8. Huang, G.B., Learned-Miller, E.: Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst (May 2014)
9. Sun, Y., Wang, X., Tang, X.: DeepID3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2014)
10. Fan, H., Yang, M., Cao, Z., Jiang, Y., Yin, Q.: Learning compact face representation: Packing a face into an int32. In: Proc. of ACM Int'l Conf. on Multimedia, ACM (2014) 933–936
11. Kemelmacher-Shlizerman, I., Seitz, S., Miller, D., Brossard, E.: The MegaFace benchmark: 1 million faces for recognition at scale. ArXiv e-prints (2015)
12. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: Proc. of IEEE Int'l Conf. on Image Proc. (ICIP). (Oct 2014)
13. Panis, G., Lanitis, A.: An overview of research activities in facial age estimation using the FG-NET aging database. In: Proc. of the European Conf. on Computer Vision (ECCV) Workshops. (2014)
14. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR). (2011)
15. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR). (June 2014)
16. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
17. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR). (June 2015)

18. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC). (2015)
19. Eastman, G.: Camera. US Patent 388850 A (1888)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), MIT Press (2012) 1097–1105
21. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems (NIPS), MIT Press (2001) 849–856
22. Belkin, M., Niyogi, P.: Semi-supervised learning on riemannian manifolds. Journal of Machine Learning **56**(1-3) (June 2004) 209–239
23. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: Proc. of Int'l Conf. on Machine Learning. (2003) 912–919
24. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schlkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems (NIPS), MIT Press (2004) 321–328