ACM MULTIMEDIA 2016 Grand Challenge
MSR Image Recognition Challenge
http://research.microsoft.com/irc/
(Latest Update: March 16, 2016)

Challenge Overview
With the success of previous MSR Image Retrieval/Recognition Challenges (MSR IRC) at IEEE ICME, ACM Multimedia 2014 and 2015, Microsoft Research is happy to announce MSR IRC at ACM MM 2016, based on real world large scale dataset, and open evaluation system.

Thanks to the advance of deep learning algorithms, great progresses have been made in visual recognition in the past several years. But, there is still a big gap from these academic innovations and practical intelligent services, due to the lack of: (1) real world large scale data with better quality for training and evaluation. (2) public platform to conduct fair, efficient evaluations and make the recognition result reproducible and accessible.

To further motivate and challenge the academic and industrial research community, Microsoft is releasing MS-Celeb-1M, a large scale real world face image dataset to public, encouraging researchers to develop the best face recognition techniques to recognize one million people entities identified from Freebase. In its V1.0 version, the dataset contains 10M celebrity face images for the top 100K celebrities, which can be used to train and evaluate both face identification and verification algorithms.

Moreover, Microsoft Research has developed Prajna Hub, an open multimedia gateway, to convert latest algorithms into online services that can be accessed by anybody, from anywhere, and make the evaluation/test results repeatable and comparable.

By participating in this challenge, you can:

- Leverage real world large scale face dataset for celebrity recognition;
- Try out your image recognition system using real world data;
- See how it compares to the rest of the community’s entries;
- Get to be a contender for ACM Multimedia 2016 Grand Challenge;

Task Description
This year we will focus on face recognition task. The contestants are asked to develop image recognition system based on, but not limited to, the datasets provided by the Challenge (as training data) to recognize 1M celebrities from their face images.

The 1M celebrities are obtained from Freebase based on their occurrence frequencies (popularities) on the web. Grounding the face recognition task to a knowledge base has many advantages. First, each people entity on Freebase is unique and clearly defined without disambiguation, making it possible to define such a large scale face recognition task. Second, each entity naturally has multiple properties (e.g. gender, date of birth, occupation), providing rich and valuable information for data collecting, cleaning, and multiple task learning.

The measurement set consists of 1000 celebrities sampled from the 1M celebrities, and for each celebrity we have manually labeled up to 20 images scrapped from commercial image search engines. But the identities of these 1000 celebrities will not be disclosed, so that the contestants cannot optimize just for these 1000 celebrities. To obtain
high recognition recall and precision rates, the contestants will have to develop a recognizer to cover as many as possible celebrities, which will be of great value to help advance the state of the art in face recognition.

A contesting system is asked to produce 5 labels for each of the test images, ordered by confidence scores. Top one and five accuracies will be evaluated against a pre-labeled image dataset, which will be used during evaluation stage.

**Dataset**

To facilitate the above face recognition task, we provide a large training dataset which covers the top 100K celebrities. This training dataset is prepared by the following steps. First, we select the top 100K entities from the 1M celebrity list in terms of their popularities. Then, we leverage public search engines to provide approximately 100 images for each celebrity, resulting in about 10M web images. Note that the dataset is mainly to facilitate the participants to quickly get started. In the contest, we do not limit the use of external data, but encourage the participants to treat data collection as part of the face recognition challenge.

Two examples are shown in Figure 1 and Figure 2 below. As shown in the figures, same celebrity may look very differently in different images. In Figure 1, Lady Gaga looks visually different due to different lighting, different poses, and heavy makeups. In Figure 2, images of Steve Jobs contain the photos when he was about 20/30 years old, as well as images when he was about 50 years old. The image at row 9, column 4 in Figure 2 is claimed to be Steve Jobs when he was in high school. Notice that the image in the right corner in Figure 2, 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”.

![Figure 1. Examples of the training images we provided for the celebrity: Lady Gaga](image-url)
As we have mentioned earlier, 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 manual 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. 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 even do not limit the use of outside training data.

Besides, the current provided training set in MS-Celeb-1M.v1 only covers about 75% of celebrities in our measurement set, which implies that the upper bound of recognition recall rate based on the provided training data cannot exceed 75%. This is mainly due to two considerations. First, being 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 all the 1M celebrities in the future. Second, we encourage the participants to treat the dataset development as one of the key problems in this challenge, and explicitly point out that external datasets are not prohibited.

For these reasons, to define a problem that is challenging enough and can be studied for years, we develop the measurement set by sampling 1000 celebrities from 1M rather than from 100K celebrities. We hope the uncovered 25% of celebrities in the measurement can encourage people label their data with entity keys in the freebase snapshot we provided and publish, so that different dataset could be easily merged to facilitate collaboration.
More details about the dataset please see the dataset document, and the dataset will be made available by 3/31/2016.

Evaluation Metric and Platform

**Guideline:** You’re encouraged to build generic system for recognizing millions of people by face. The 1M celebrity names and about 10M face images with labels will be provided to the participants for data filtering and training. A development data set, which contains several hundreds of face images and ground truth labels will be provided to the participants for self-evaluations and verifications. Please note that above datasets are all optional to be used. That is, systems that based on MS-Celeb-1M and/or any other private/public datasets will all be evaluated for final award (as different tracks, if necessary), as long as the participants describe the datasets they have used.

**Evaluation Metric:** To match with real scenarios, we will measure the recognition recall at a given precision 90% (or 99% if there are a sufficient number of teams who can achieve this level of precision). That is, for N images in the measurement set, if an algorithm recognizes M images, among which C images are correct, we will calculate precision and recall as:

\[
\text{precision} = \frac{C}{M} \\
\text{recall} = \frac{M}{N}
\]

By varying the recognition threshold, we can determine the recall when the precision is at 90%. Note that we also add distractor images to the measurement set. This will increase the difficulty of achieving a high precision, but is much closer to real scenarios.

**Evaluation Platform:** An open multimedia hub, Prajna Hub, will be used for the evaluation, which will turn your recognition program to a cloud service, so that your algorithm can be evaluated remotely. Similar methodology has been used in the last several IRCs and it was well-received. This time, we made it even easier, with extra bonus including:

- Your recognizer will be readily accessible by public users, e.g. web pages, mobile apps. But the core recognition algorithm will still be running on your own machine/clusters (or any other public clusters if preferred), so that you always have full controls;

- Sample codes for web/phone apps will also be available through open source, so that your recognition algorithms can be used across devices (PC/Tablet/Phone) and platforms (Windows Phone, Android, iOS). I.e., you will have a mobile app to demonstrate your face recognizer, but you won’t need to write mobile app codes or just need to make simple modifications.

- Sample codes will be provided to help participant to convert your existing recognition algorithms to a cloud service, which can be accessed from anywhere in the world, with load balance and geo-redundancy;

The recognizer can run on either Windows or Linux platform.

**Participation**

The Challenge is a team-based contest. Each team can have one or more members, and an individual can be a member of multiple teams. No two teams, however, can have more than 1/2 shared members. The team membership must be finalized and submitted to the organizer prior to the Final Challenge starting date.

---

1 By design, the measurement set and the development set should have the same format, and similar distribution.
At the end of the Final Challenge, all entries will be ranked based on the metrics described above. The top three teams will receive award certificates. At the same time, all accepted submissions are qualified for the conference’s grand challenge award competition.

Paper Submission
Please follow the guideline of ACM MULTIMEDIA 2016 Grand Challenge for the corresponding paper submission.

Detailed Timeline (Tentative)
• March 14, 2016: details about evaluation announced/delivered
• March 31, 2016: MS-Celeb-1M.v1 ready for download
• June 7, 2016: Dry run starts (trial requests sent to participants)
• June 13, 2016: Evaluation starts (evaluation requests start at 8:00am PDT)
• June 17, 2016: Evaluation ends (5:00pm PDT)
• June 23, 2016: Evaluation results announced.
• July 6, 2016: Grand Challenge Paper and Data Submission deadline
• July 29, 2016: Notification of acceptance
• August 3, 2016, Camera-ready submission deadline

More information
• Research paper about the dataset: "MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World"

Challenge Contacts
Questions related to this challenge should be directed to:

Yuxiao Hu (yuxiao.hu@microsoft.com), Microsoft Research
Lei Zhang (leizhang@microsoft.com), Microsoft Research
Shiguang Shan (sshan@ict.ac.cn), Chinese Academy of Sciences