Face Recognition via Active Annotation and Learning

Hao Ye[1]*, Weiyuan Shao[1]*, Hong Wang[1]*, Jianqi Ma[2],
Li Wang[2], Yingbin Zheng[1]†, Xiangyang Xue[2]
[1]Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China
[2]School of Computer Science, Fudan University, Shanghai, China

ABSTRACT

In this paper, we introduce an active annotation and learning framework for the face recognition task. Starting with an initial label deficient face image training set, we iteratively train a deep neural network and use this model to choose the examples for further manual annotation. We follow the active learning strategy and derive the Value of Information criterion to actively select candidate annotation images. During these iterations, the deep neural network is incrementally updated. Experimental results conducted on LFW benchmark and MS-Celeb-1M challenge demonstrate the effectiveness of our proposed framework.

Keywords

Face recognition; active image annotation; deep learning; MSR image recognition challenge

1. INTRODUCTION

With the development of deep learning approaches, face recognition system based on neural networks achieve great improvement on performance as well as the recognition robustness. According to recent progress of deep network applications on computer vision tasks, deeper networks can boost the performance [1, 20, 23, 6]. In order to avoid overfitting, larger amount of training images are usually needed. Moreover, precise annotations of the training images can carry more supervised information and benefit the results of deep networks. Particularly for face recognition tasks, information such as face localization, landmark points, and pose may influence the detection and recognition performance. Therefore, constructing a large-scale fine-annotated face dataset and building face models upon it become very important steps towards a successful face recognition system.

*These authors contributed equally to this work.
†Corresponding author. E-mail: zhengyb@sari.ac.cn

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2984059

Recently, many public datasets for face recognition are released and promote a significant amount of progress for face detection and recognition tasks. For instance, the CASIA-WebFace dataset [26] contains more than 400,000 images and 10,000 celebrities, VGG Face dataset [16] is with 2.6 million images and 2,600 celebrities, and MS-Celeb-1M dataset [5] has more than 10 million images and 100,000 celebrities. These datasets were collected from the search engines such as Bing and Google. Due to the massive quantities, the images are usually without further filtering and preprocessing. The influence of these insufficient preprocessing includes two aspects. Firstly, the noisy dataset with incorrect annotations and low quality images may imperil the correctness of model training and reduce the accuracy for recognition. On the other hand, multimodal information such as landmark points and regions are proved to be an important cue [3]; however, the public datasets lack of these kind of annotations.

Manually annotation of a large dataset with detail information such as celebrity label, face location, landmark points and other attributes are infeasible. Various data augmentation approaches are designed to enrich the information and overcome the deficiency of annotated data. [15] employed 3-D alignments to get images with extra face pose and landmark points information. However, it is difficult to extend the celebrity popularity, which leads to the lack of diversity of training samples. Another strategy is to combine several small or partial annotated datasets through multi-task strategy to simulate the larger dataset (e.g. [27]). Although it can alleviate the shortage of data and annotation, this method lacks flexibility to adapt itself to new dataset since each task is designed based on specific dataset.

In this paper, we propose an active annotation and learning framework to handle the training of face recognition models on large-scale dataset with noisy data and insufficient annotations. We design an active sampling strategy and employ the iterative training methods to update the face model. The advantages of active annotation system are in two folds. First, the framework is effective and efficient in sampling hard negative examples which saves costs of annotation. Comparing with the large amount of entire dataset, we only annotate a small subset of the whole dataset and the recognition system can also achieve high performance. Second, the framework is quite flexible - not only new images but also new attributes can be annotated, as the framework is incremental and new annotation tasks can be added at the end of each iteration. The experimental results on face

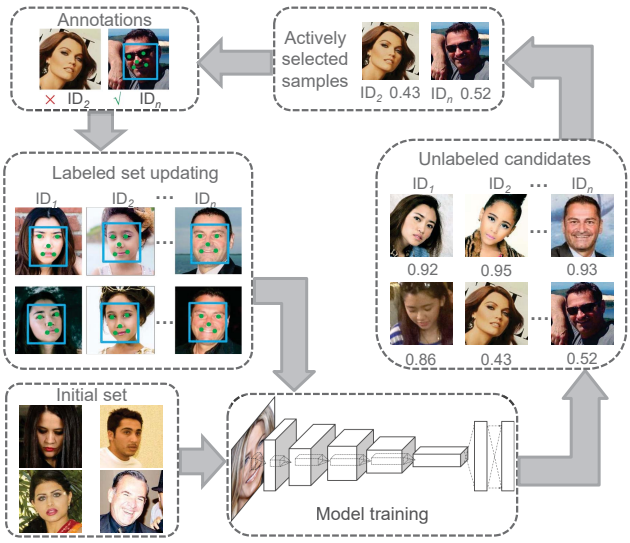


Figure 1: Our face annotation and recognition framework.

recognition benchmarks demonstrate the effectiveness and applicability of our framework.

The remainder of this paper is organized as follows. Section 2 reviews the related work on task of image annotation and face recognition. Section 3 describes our proposed framework. In Section 4, we demonstrate the quantitative study on two face recognition benchmarks. And finally, we conclude our work in Section 5.

2. RELATED WORKS

Constructing a large-scale face dataset requires a lot of time and efforts. Google has built a massive dataset which includes 200 million face identities and 800 million image faces [18], but it is a private dataset. [26] proposed a semi-automatical way to collect 494,414 face images from internet and build the CASIA-WebFace dataset. [12] released the MegaFace dataset, which includes 1 million images and 690,000 identities. Very recently, Microsoft Research provided a face dataset involving 100,000 identities and 10 million images, which is the largest open face dataset to the best of our knowledge [5].

Active learning has gained attention in the field of large-scale image annotation. [7] employed the multi-label active learning engine to automatically select and manually annotate a set of unlabeled sample-label pairs, and then the classifier is updated by taking these newly annotated pairs. [17] integrated machine and human as a loop for object detection annotation and validated on ILSVRC 2014 detection task empirically. [2] proposed an active learning algorithm to measure the sample uncertainty where the density and diversity of the sample were determined simultaneously. [21] introduced a criterion that propagates the human correlation on selected samples over a gradually-augmented graph. [19] used an automatic selection method to automatically choose a small number of hard negative examples which made the training effective and efficient.

3. FRAMEWORK

We now elaborate the construction of our framework, which consists of initial model training, active sampling, and iterative model training. Figure 1 illustrates the architecture of the framework and the details of each phase will be described in the following subsections.

Figure 2: Examples of the noisy images. Left: low resolution; Middle: blur; Right: animation.

3.1 Data Preprocessing

Before we dive into the framework, a data preprocessing step is applied on the original dataset. As the amount of face image increases, some noisy images are inevitably mixed in the dataset, such as low-resolution images, blur images and animation images; examples of these images are illustrated in Figure 2. Although the deep network model has the ability of fault tolerance to a certain extent, the accuracy of the recognition system is affected by these noisy data. Thus, filtering the noisy data out of dataset is necessary and considered as the preprocessing step. Therefore we implement noisy detection tools based on OpenCV [9] to detect these kinds of noisy images.

3.2 Initial Model Training

Our face recognition pipeline includes face detection and landmark point localization, face alignment and transformations, and recognition based on deep network. Considering our main purpose is to inspect the effectiveness of active annotation scheme on recognition, we take the off-the-shelf toolkit dlib [13] for face detection and landmark point localization, as most existing face datasets lack such annotations. Images with detected faces are considered as the initial model training set.

We perform the alignment and transformation operations when corresponding facial landmark points in each face are obtained. Considering the time complexity, we select in-plane 2D alignment method which only use 5 landmark points (i.e., nose, left eye, right eye, left mouth, and right mouth) to align the face image. Then we align and transform the faces based on the landmark points to weaken the influence from the pose variation and augment the versatility of faces.

The deep network architecture of face recognition model is shown in Figure 3. The input of the network is $X=\{(x_1, y_1), (x_2,y_2), ..., (x_n, y_n)\}$, where x_i is the face image after align transformation and y_i is the corresponding identity. We follow the network architecture of lighted CNN [25], which consists of 9 convolutional layers (conv) and 2 fully connected layers (fc). Maxout [4] is used as the active function, and cross entropy loss as the supervisory signal.

3.3 Active Sampling

With the trained deep models, the next step is to actively sample images for annotation. The goal of active sampling is to minimize human interventions. We extend the value of information (VOI) strategy proposed in [11] and derived a value function to choose the instance for annotation.

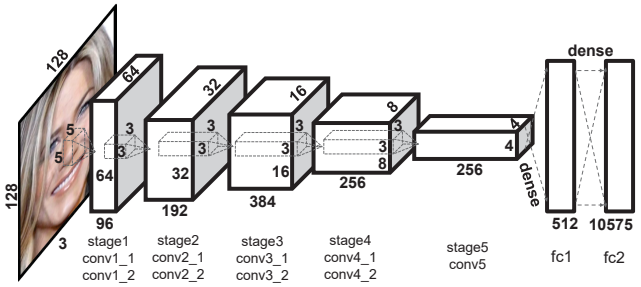


Figure 3: Network design for face recognition.

Each dataset image is with an initial identity, as all of them are gathered from the web using the identity name as keyword. If the image is classified into the same identity, we consider it as a positive instance; otherwise it is negative. The dataset includes unlabeled set X_U and labeled set X_L; both of them are composed of positive instances X_p and negative instances X_n. The risk associated with misclassifying a positive example as negative is denoted by r_p, and r_n for misclassifying a negative. The risk associated with the annotated examples is:

$$Risk(X_L) = \sum_{x_i \in X_p} r_p(1 - p(x_i)) + \sum_{x_i \in X_n} r_n p(x_i), \quad (1)$$

where $p(x)$ denotes the probability of positive prediction, here we use the softmax of the confidence from the network. The corresponding risk for unlabeled examples is:

$$Risk(X_U) = \sum_{x_i \in X_U} [r_p(1 - p(x_i)\Pr(y_i|x_i) + r_n p(x_i)(1 - \Pr(y_i|x_i))], \quad (2)$$

where y_i is the true identity label. We approximate the probability as $\Pr(y_i|x_i) \approx p(x_i)$ which lead to a simplifier equation:

$$Risk(X_U) = \sum_{x_i \in X_U} (r_p + r_u)(1 - p(x_i))p(x_i) \quad (3)$$

The total cost associate with the data is the total risk in addition with the cost of annotation,

$$T(X_L, X_U) = Risk(X_L) + Risk(X_U) + \mathcal{C}(X_p) + \mathcal{C}(X_n), \quad (4)$$

where $\mathcal{C}$ is the cost of obtaining the annotations. We ignored $\mathcal{C}$ in our experiments under the assumption that the cost of obtaining annotations for each example is the same. The expected risk of candidate samples for annotation can be calculated by:

$$\begin{aligned} VOI(S) &= T(X_L, X_U) - T(X_L \cup S, X_U \setminus S) \\ &= Risk(X_L) + Risk(X_U) \\ &\quad - (Risk(X_L \cup S) + Risk(X_U \setminus S)) \end{aligned} \quad (5)$$

where S denotes the candidate sampling examples for manual annotation. According the above strategy, the high value indicates large gains and thus our strategy is to choose candidates S with highest VOI for labeling.

3.4 Iterative Training Scheme

The samples selected are then used for the manual annotation; the annotation of both bounding box and 5 landmark points are required. If the annotator confirms an image without its original identity, this image will be removed from the updating set, meanwhile, the positive images with their manually annotation are added into the updating dataset. The deep network is incrementally updated, using the fine-tuning strategy. With the precision level of the annotation, the number of active sampling is gradually decreasing whereas the recognition becomes more robust and accurate. We stop the iterative training when the VOI of the unlabeled images are below a threshold and thus no new samples are selected.

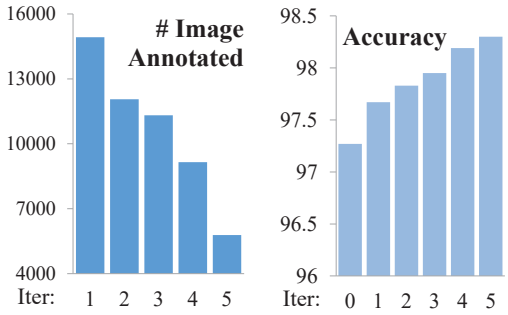


Figure 4: Active sampling image number and performance of LFW after each iteration.

Approach	#Nets	Accuracy
WebFace [26]	1	97.73%
Lightened CNN [25]	1	98.13%
Face Search [24]	7	98.23%
MM-DFR-JB [3]	1	98.43%
Ours	1	98.30%

Table 1: Comparison with previous works using CASIA-WebFace for training.

3.5 Application of Face Models

At any stage of our framework, the deep network can be applied for the face recognition task. As mentioned in previous works (e.g. [22]), if the network can discriminate large number of individuals, the feature from the fully connected layer is a good representation in face recognition task. For an input image, after face detection and alignment, a 512-dimensional deep feature from the response of fc1 is extracted and used as the representation. Finally, we use cosine distance as the similarity metric on the representation and k-Nearest Neighbor classifier for face recognition.

4. EVALUATION

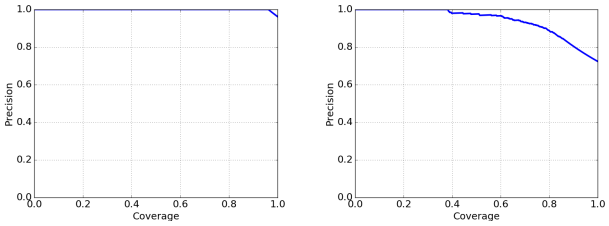
Our framework is based on the annotation and training of the CASIA-WebFace dataset [26]. It contains about 494,414 images with 10,575 individuals, and is a large public available face recognition dataset. The deep network is trained on Caffe [10] and the initial network is trained on an NVIDIA GTX Titan X GPU for about 60 hours.

4.1 LFW

We first evaluate our method in the Labeled Faces in the Wild (LFW) benchmark [8], which is the most popular benchmark for face recognition task. The dataset contains 13,233 face images with 5,749 individuals, and our ex-

(a) Random set (b) Hard set

Figure 5: Examples of dry run images in MS-Celeb-1M.



(a) Random set (b) Hard set

Figure 6: The Precision-Coverage curve of our system for the dry run.

periments follow the unrestricted settings which allow using external data in the training period.

In order to test the effectiveness of active sampling, the deep models after each iteration are saved and the corresponding features of the LFW images are extracted and used for the benchmark. Figure 4 illustrates the number of active sampling images from CASIA-WebFace and accuracy of LFW after each iteration. Comparing with the image number of the whole CASIA-WebFace dataset, only about 10% of them are actively sampled for annotation. We also notice that the sampling images in each iteration reduce, while the performance of our deep model improves sustainingly. Table 1 compares our method with previous approaches which are also trained using CASIA-WebFace. The performance of our model reaches the same magnitude of these methods.

4.2 MS-Celeb-1M

We also participate in MSR Image Recognition 2016 Challenge[1]. MS-Celeb-1M [5] is 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 10 million celebrity face images for the top 100,000 celebrities, which can be used to train and evaluate both face identification and verification algorithms. And the dataset provides approximately 100 images for each celebrity, resulting in about 10 million web images. Current dataset only covers 75% of celebrities, which implies that the upper bound of recognition recall rate cannot exceed 75%.

The measurement set samples 1000 celebrities from the 1 million set, which indicates that the remaining 25% celebrities may also appear. To match with real scenarios, the evaluation system will measure the recognition recall at a given precision 95%. For N images in the measurement set, if an algorithm recognizes M images, among which C images are correct, the precision and coverage will be calculated as: $Precision = \frac{C}{M}$, $Coverage = \frac{M}{N}$.

[1]https://www.microsoft.com/en-us/research/project/ms-celeb-1m-challenge-recognizing-one-million-celebrities-real-world/

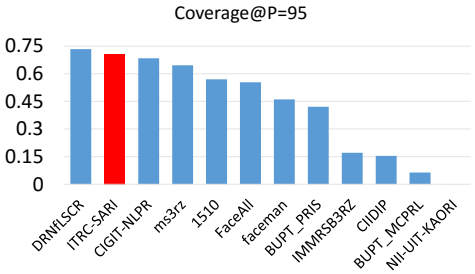


Figure 7: The evaluation performance of our system in MS-Celeb-1M benchmark.

There are two tracks in the Challenge: Random set and Hard set; examples of both sets are shown in Figure 5. the face images in Hard set are with big variations, while the test images in random set are randomly selected and highly likely to be covered by the MS-Celeb-1M training data. Figure 6 is the Precision-Coverage curve of our face model on the dry run. Our approach perform better on Random set, as many faces in Hard set (e.g. profile images) cannot be detected using dlib detector [13] and thus failed to return results.

We also compare our model with other teams on Random set and Figure illustrates the comparison. Our model reaches the coverage 70.7% when Precision=95% and ranks $2rd$ in the Random set competition. Note that during the training phase of the evaluation, we don't use the MS-Celeb-1M training data; the same pipeline and model as LFW are employed. We believe that the performance of our model can be improved after the applying of active framework on MS-Celeb-1M, and we would like to examine in the future.

5. CONCLUSION

In this paper, we have introduced a human-in-the-loop framework for face recognition. After training a deep network from initial face dataset and applying the feature from the network, all the face images have a confidence for its associated celebrity label. The VOI strategy is employed to actively select images and these images are illustrated to the annotator. After each batch of request, the deep network is incrementally updated. The whole framework runs iteratively, and the deep model can be used for the face recognition task at any stage. Experimental comparisons with previous works on LFW and MS-Celeb-1M show the effectiveness of proposed active annotation and learning framework on face recognition task.

Our future work will test our framework on the larger dataset, i.e. the training dataset of MS-Celeb-1M Challenge; we would like to further analysis the relation between the face recognition performance and the manual annotation effort. Another potential direction may be replace our current face detector with state-of-the-art face detection methods such as cascaded CNN [14] and build an end-to-end active and learning system base on deep networks.

Acknowledgment

This work was supported in part by grants from Natural Science Foundation of China (No. 61572138) and Science and Technology Commission of Shanghai Municipality (No. 16511104802).

6. REFERENCES

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[2] M. Chatterjee and A. Leuski. An active learning based approach for effective video annotation and retrieval. *NIPS*, 2015.

[3] C. Ding and D. Tao. Robust face recognition via multimodal deep face representation. *IEEE Transaction on Multimedia*, 17(11):2049–2058, 2015.

[4] I. J. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *ICML*, 2013.

[5] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large scale face recognition. In *ECCV*, 2016.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[7] X.-S. Hua and G.-J. Qi. Online multi-label active annotation: Towards large-scale content-based video search. In *ACM Multimedia*, pages 141–150, 2008.

[8] 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.

[9] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014.

[11] A. Kapoor, E. Horvitz, and S. Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *IJCAI*, 2007.

[12] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016.

[13] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[14] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *CVPR*, pages 5325–5334, 2015.

[15] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? *arXiv preprint arXiv:1603.07057*, 2016.

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[17] O. Russakovsky, L.-J. Li, and L. Fei-Fei. Best of both worlds: Human-machine collaboration for object annotation. In *CVPR*, 2015.

[18] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[19] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[21] H. Su, Z. Yin, T. Kanade, and S. Huh. Active sample selection and correction propagation on a gradually-augmented graph. In *CVPR*, 2015.

[22] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, pages 1891–1898, 2014.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

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

[25] X. Wu, R. He, and Z. Sun. A lightened cnn for deep face representation. *arXiv preprint arXiv:1511.02683*, 2015.

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

[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *arXiv preprint arXiv:1604.02878*, 2016.