A way to improve precision of face recognition in SIPP without retrain of the deep neural network model

Xihua.Li
lixihua9@126.com

Abstract

Although face recognition has been improved much as the development of Deep Neural Networks, SIPP(Single Image Per Person) problem in face recognition has not been better solved. In this paper, multiple methods will be introduced to improve the precision of SIPP face recognition without retrain of the DNN model. First, a modified SVD based method will be introduced to get more face images of one person in order to get more intra-class variations. Second, some more tricks will be introduced to help get the most similar person ID in a complex dataset, and some theoretical explain included to prove of why our tricks effective. Third, we would like to emphasize, no need to retrain of the DNN model and this would be easy to be extended in many applications without much efforts. We do some practical testing in competition of Msceleb challenge-2 2017 which was hold by Microsoft Research and finally we rank top-10. Great improvement of coverage from 13.39% to 19.25%, 29.94%, 42.11%, 47.52% at precision P99(99%) would be shown in this paper.

1. Introduction

Since 2012, Deep Neural Networks has been utilized in almost every aspects of computer vision, such as image classification and recognition, object detection, tracking and recognition, OCR in natural images, face recognition, saliency detection, 3D action recognition, image segmentation, super-resolution, image creation, content-base-image-retrieval, medical image diagnosis, and so on. At the same time, deep neural networks itself has also improved much, from AlexNet to GoogleNet, VGG, ResNet, Inception, Inception-ResNet[1], which focus on mainly two aspects. First, to reduce the params of deep neural networks. Second, to improve the feature extraction and representation ability of deep neural networks. The destination is to find a neural networks which has the best representation of what's in the picture with less computing.

Face Recognition is almost the most hot topic in computer vision and has now been put into use in many practical applications. At early stage, Fisher face and Eigen face[2] were used to regard face recognition as a problem of finding a suitable projection that can mostly maximum inter-class variations and minimum intra-class variations. In middle stage, face landmarks were detected and local features such as LBP are extracted to have a better representation of face with contextual information included. Nowadays, almost every top face recognize algorithms are deep neural networks based. With different kinds of loss function design, deep neural networks based face recognize methods has improved much and now 99.8% precision has been reached on LFW benchmark which has already exceed human beings[3].

Different kinds of loss function is also a way for how us human beings to treat what kind of machine learning problem face recognition is. At the beginning, face recognition was formulated as a classification problem, but when face IDs increase sharply, classification precision decreased. Then, verification based loss function occur, deep neural networks was used to learn a representation that could verification whether the two faces belongs to a same person or not. DeepID, DeepID2, DeepID3[4] perhaps is the most successful face recognition algorithm that profit from the design of loss function.

Different kinds of networks and loss function were essentially used to learn a powerful representation of human face with lower intra-class variations and higher inter-class variations. But if we already learnt this representation, is there any way that we could do to improve the face recognition precision without much efforts. Under this consideration, large scale face recognition problem is a way for us to search for the most nearest points in high dimensions feature space when given a particular face image feature representation vector. In this paper, we take Msceleb challenge-2 2017 hold by Microsoft Research as background and do every evaluation on dataset of the challenge, some tricks were proposed to improve the precision of face recognition without retrain of the DNN models.

First, in section 2 a face augmentation method based on SVD would be introduced. Second, multiple tricks would be introduced with different kinds of logic behinds and the-

oretical explain followed in section 3. Third, in section 4, we do a comprehensive evaluation on Msceleb challenge-2 2017 and also, params used in the competition would be listed. Finally, conclusions and future work directions introduced in section 5.

2. Msceleb challenge-2 2017 and face augmentation

2.1. Msceleb challenge-2 2017

Msceleb face recognition challenge[5][6] was hold by Microsoft Research which was named the "World Cup" for face recognition, attract many researchers[1]. Challenge-2 was newly started, mainly focus on SIPP problem and also emphasize the generalization ability of the face recognition algorithm, which is almost the real-world scenarios. Challenge-2 is called Low-Shot Learning or Know you at One Glance. Here we have a brief introduction.

In challenge-2, we investigate the problem of low-shot face recognition, with the goal to build a large-scale face recognizer capable of recognizing a substantial number of individuals with high precision and high recall. We create a benchmark dataset consisting of 21,000 persons each with 50-100 images of high accuracy (>99%). We divide this dataset into the following two sets: **Base set**, there are 20,000 persons in the base set. Each person has 50-100 images for training, and about 5 images for testing. **Novel set**, there are 1,000 persons in the novel set. Each person has 1-5 images for training, and 20 images for testing.

Our goal is to study when tens of images are given for each person in the base set while only one to five images are given for each person in the novel set, how to develop an algorithm to recognize the persons in both the data sets.

Our measurement set contains a mixture of test images from both the base set and the novel set. We mainly focus on the classification performance with the test images in the novel set to evaluate how well the computer can learn novel visual concepts with limited number of training samples, while also monitor the performance on the base set to ensure that the performance gain on the novel set which is not obtained by sacrificing the performance on the base set. A contesting system is asked to produce at least one prediction label with a confidence score per test image. To match with real scenarios, we measure the recognition coverage at a given precision 99%. 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 coverage as:

$$precision = C/M \tag{1}$$

$$coverage = M/N \tag{2}$$

[1]http://www.msceleb.org/

Figure 1: Examples of SVD augmented faces. Left to right: 100%, 95%, 90% percentage of energy reserved. Top to bottom: noise added, blur, noise plus blur effects occurs for different faces.

By varying the recognition confidence threshold, we can determine the coverage when the precision is at 99%. We rank the methods according to the coverage at the 99% precision with the test images in the novel set, while monitor the performance on the base set.

2.2. SVD based face augmentaion

In this section, we propose a modified SVD-based methods to do face augmentation for novel set faces. By using SVD[7], we decompose the face image into two complementary parts: the first part in constructed by the SVD basis images associated with several largest singular values, and the second part is constructed by the other low-energy basis images. This first part preserves most of the energy of an image and reflects the general appearance of the image. The second part is the difference between the original image and the first part, and it can reflect, to some extent, the variations of the same class face images.

Given a aligned face image $A \in R^{m*n}$ and suppose $m \geqslant n$, we have the following expression according to SVD.

$$A = \sum_{i=1}^{n} \sigma_i \cdot \mu_i \cdot \upsilon_i \tag{3}$$

where μ_i are ith column of $U \in R^{m*m}$ and $V \in R^{n*n}$, respectively. U and V are composed of the eigenvectors of AA^T and A^TA, respectively, σ_i is the singular values of image A and we let $\sigma_1 \geqslant \sigma_2 \geqslant ... \geqslant \sigma_n$.

In our experiment, we try to reserve 95%, 90%, 85%, 80%, 75% energy of the eigenvalues at first. After some more evaluations, we reserve 95%, 90%, 85%, in that way,

one face image become 4 in total. In general, SVD faces were always generated on grayed image. In this paper, we do SVD face on each of RGB channel and merge them up, for each channel, same or different percentage of energy reserving could be used. Augmented face images could be seen in figure 1.

Here we have brief introduction about the baseline and evaluations methods. First, a Dlib[2] based face detection, face alignment and face feature extractor method was used in whole of the evaluation. Dlib deep neural network face recognition tools has a 99.38% accuracy on the standard LFW face recognition benchmark, which is comparable to other state-of-the-art methods for face recognition until early of 2017. It's essentially a version of the ResNet-34 network from the paper Deep Residual Learning for Image Recognition by He, Zhang, Ren, and Sun with a few layers removed and the number of filters per layer reduced by half. The network was trained from scratch on a dataset of about 3 million faces. Second, for Base set there are 20,000 persons and each has 50-100 images, for Novel set there are 1,000 persons each has only one image. We do face detection, alignment and face feature extraction on Base set and Novel set and finally got 1,169,166 face feature vectors for 1,169,166 face images which belongs to 21,000 persons. Note please, for Novel set, we do SVD on each face images and got 3 degrade face images for correspondence person, we also extract feature vectors on degraded face images with dilb face feature extractor, and more 3,000(=3*1,000) face feature vectors we got which totally became 1,172,166 face feature vectors. Notations, $X_{i,j}$ is base set feature vector where $i \in [1, 20000]$ and $j \in [50, 100]$, $Y_{i,j}$ is novel set feature vector where $i \in [1, 1000]$ and $j \in [1, 4]$, where j=1 represents the original face feature vector and j=2,3,4 represents SVD degraded face feature vectors for novel set.

2.3. SVD based method evalution

For test set, we only focus on novel set persons, but search for the person id in all Base set and Novel set. After dlib face feature extractor we finally got 4,899 face feature vectors although totally there are 5,000 face images in test set(some images may not detect any face with dlib).

Easiest, we just do a brute force search over all face feature vectors of 21,000 person in base set and novel set(SVD degrade features are not included), search for the nearest face in totally 1,169,166 face images. And then we search for the most similarity person id for test set in way of compare with all feature vectors, similarity score defined bellow, and maximum similarity score return the particular person id. After a truly time consuming search, the final result is 13.39%@P99, 33.41%@P97, 56.87%@P95.

$$SimilarityScore = 1.0/(1.0 + dist(X, Y)) \quad (4)$$

[2]http://vis-www.cs.umass.edu/lfw/results.html

Methods	P99	P97	P95
Base	13.39%	33.41%	56.87%
Trick1	19.25%	39.64%	72.69%
Trick2	29.94%	92.65%	100%
Trick3	42.11%	93.55%	100%
Trick4	47.52%	94.2%	100%

Table 1: Coverage for all tricks we used. Base represent brute force search over no SVD added feature vectors.

$$dist(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (5)$$

Then, when SVD degrade features were included, and the result is 19.25%@P99, 39.64%@P97, 72.69%@P95, which emphasize that SVD is truly useful for SIPP face recognition. Theoretically, when SVD face feature vectors included, it sames like that we added some intra-class variations for person in novel set, or more precisely, we have expanded high dimension feature space for novel set face image from on point to a hypersphere around the point. Showing in figure 1, in right circle, expand from blue point to blue circle.

It's really not a good result and we began to add more tricks in search and we get the evaluation result in table 1.

As we could see, coverage has improved to 47.52% after 4 tricks added on Base method, and Coverage reach to 100%(P95) when only 2 tricks had added, that is to say, when a precision of 95% ensured, we can recognize all the people in test set. Next, we will introduce each of the tricks and its theoretical explain followed.

3. Tricks for search the person ID

In this paper, we do not trying to training a deep neural networks that can have a good representation of each person with minimum intra-class variations and maximum inter-class variations. We focus on the search strategy for finding the corresponding person ID. That means when a face feature extractor fixed, we still have many methods to improve the coverage and precision in SIPP face recognition problem.

3.1. Trick1: search over all feature vectors with SVD face added

As we have introduced in section 2.3, SVD face with 95%, 90%, 85% energy reserved were included in the search space. On on hand, it's a way for us to added intra-class variations for persons in novel set. On the other hand, SVD added faces is also a way for us to expanding the representation of a person from one point to one hyperspace surrounding the point, although the hyperspace is not in the

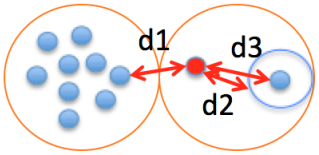

Figure 2: Explaining of why SVD face useful. Left circle: face from base set which has many face feature vectors full-filed corresponding hyperspace. Right circle: face from novel set which has only on face feature vector and expanded after SVD faces added. After expanded, d2 could be found in order not to be mis-classified.

middle of the person feature space. This situation explained in Figure 2.

After a heavy analysis on the evaluation, we found it's easy for us to find a person whose expression, light conditions, point of view and so on is quite similar to the search face but comes from another person ID. In fact, this situation can also to be explained with Figure 2, although the hyperspace of novel set person has been expanded with SVD faces, but the expanding is limited because of SVD faces could only add variations in one way or in other words, the expand hyperspace is limited. How about we just utilize of mean feature vector to represent a person.

3.2. Trick2: search over mean feature vectors

For trick 2, we compute the mean feature vectors for each person in base set and novel set, and then the search would just be done over all the mean vectors. That means, only the main information of a person should be utilized, in order not to be misguided by variations such as expressions, light conditions, point of view which is not the essential difference between persons. But there is still drawbacks, for person in novel set, the mean feature vector is not in the middle of the high dimensions hyperspace of the person, but just located on the only one face feature vector we have. As the only on face feature vector is a little far from the middle of hyperspace of the person, it will always lead to not very high similarity score although its within the same person.

After trick 2, it's clear that we've found a effective way to search for person ID over base set and novel set, as a coverage of 100% has reached for precision 95%(P95). So we keep mean search all the way with all tricks we have next. Is brute force search does not work at all? We do not think so. In fact, brute force search can help to get a more distinguishable similarity score when two faces belongs to one person. So we do a combination of mean search and brute force search.

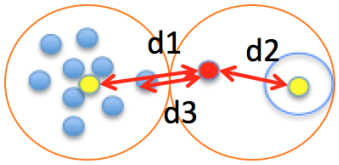

Figure 3: Explaining of why mean face feature vector useful. Although d3 is the most similar one, but when we just search over mean vector, only d1 would be founded, which would not lead to a mis-classified.

PersonId	SameIdNot	SimilarityScore
0	1	0.92
1	1	0.91
2	1	0.90
3	1	0.89
4	1	0.88
5	-1	0.87
6	1	0.86
7	-1	0.85
8	1	0.84
9	-1	0.83

Table 2: Explain why a distinguishable similarity score is important. If we want to have a precision 100%, the threshold will be 0.88 while coverage is 50%. As person ID 6 and 8 has a low similarity score of same person respectively, If we can assign them a more distinguishable similarity score bigger than 0.88, coverage would be increased.

3.3. Trick3: combination of mean search and brute force search

As explained in trick 2, mean search is a way for us to ensure a minimum coverage in challenge-2, and brute force search is a way for us to get a more distinguishable similarity score which could increase the coverage a further step at a high precision(P99) required. In this part, an example why a more distinguishable similarity score is help and how we do will be introduced.

In table 2, as the explaining, if we can assign a more distinguishable similarity score for which two face came from the same person with a high confidence, coverage at a particular precision could be improved. If we want to get a higher similarity score, a brute force search could help in the way of d2 in figure 2. Also, the same with situation in figure 3, d3 would not be choose because of a mis-classification would occur. After a heavily analysis on evaluations, if we can find a face feature vector with a higher similarity score

$$SimilarityScore, Id = \begin{cases} max(s1, s2), id1 & if(id1 = id2) \\ \begin{cases} s2, id2 & if(s2 > s1 + T) \\ s1, id1 & if(s2 < s1 - T) \\ min(s1, s2), id1 & if(s2 \in [s1 - T, s1 + T]) \end{cases} & if(id1 \neq id2) \end{cases} \quad (6)$$

which exceed the similarity score of mean feature vector with a threshold T, that would be a important indication of same person, and if not, they always came from two persons. Based on this observation, we do a combination of mean search and brute force search within a trick way. Suppose after mean search, we found a most similar similarity score s1 and the corresponding person ID id1, after a brute force search, another most similar similarity score s2 and the corresponding person ID id2. The combination was done with rules explained in equation 6.

If mean search and brute force search find the same person ID, we choose the maximum score between mean search and brute force search because we have a high confidence of a correct search. If mean search and brute force search indicate different person IDs, we choose the one who exceed the other with a score more than a threshold T. If the limitation of threshold T could not satisfied, we choose the minimum score between s1 and s2, and the mean search id for mean search has a higher confidence than brute force search in some extent. In this paper, threshold T is set to 0.03 for all evaluations.

3.4. Trick4: combination of mean search and nearest neighbor search

On one hand, a brute force search method is extremely time consuming which could not been used in realtime applications. On the other hand, if noise exists in feature vectors, a brute force search method tends to mis-match with the noise as a higher similarity score may always been obtained at a constant probability. In order to handle the two weak points explained here, a nearest neighbor search method was used to do the combination with mean search. As nearest neighbor search always used in CBIR, K-D tress, structured K-means, LSH[8] are most commonly used methods. In this paper, a LSH based nearest neighbor search method was used and we set the target precision of search at 0.98 to ensure a good precision. As listed in table 1, we finally reach a coverage of 47.54% at P99 and a coverage of 94.2% at P97.

4. Results and Evaluations

In this part, we will show more details for each evaluations in challenge-2. As shown in figure4, precision-coverage curves was drawn. First, on the right side, when required precision is only 0.87, all 5 methods could reach a coverage of 100%. Second, for trick2, trick3, trick4, the final precision for all person in novel set is 95%, which means the total number of mis-classified face images is same with the 3 methods. But, why it has a different coverage when precision is 99%, the only explanation is that there is slightly different confidence score for some images, and that's why we are looking for a more reasonable way to assign a similarity score. Third, on the left side, base0 and trick 1 has a sharply drop at the beginning, that means we have assign a bigger similarity score for mis-matched faces or smaller similarity score for matched faces. After mean vector search added, this phenomenon disappeared, which also demonstrate the efficient of mean vector search.

Coverage over P97 and P95 are shown in figure 5 and figure 6. As coverage reaches 100% when precision requirement is 95%. We could also result in another conclusion, if a better face feature extractor would be obtained, which means the totally precision is higher than 95%. In this way, we surly could get a higher coverage over P99.

5. Conclusions

This paper do not aimed at design or training a powerful neural networks for SIPP face recognition problem, but pay more attention to proposing some search tricks which would significantly increase the coverage under a given precision requirement with logic behind. First, a modified SVD based face image generate method was used to produce more intra-class variations. Second, 4 tricks were proposed for more precise and effective search of correct person ID and some theoretical explain followed. Third, we would like to emphasize, no need to retrain of the DNN model and this would be easy to be extended in many applications without much efforts. Coverage under P99 improved from 13.39% to 47.52% without retrain of model and much computing, and if we lower the precision requirement to P95, 100% coverage obtained. In the future, one on hand, we will focus on produce more intra-class variations for SIPP faces, with the way such as generative adversarial networks(GAN), one the other hand, more search tricks could be evaluate for effective search with higher precision. Of course, we will try to design more powerful deep neural networks for a better face feature extractor to improve the performance of the whole system.

References

[1] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the im-

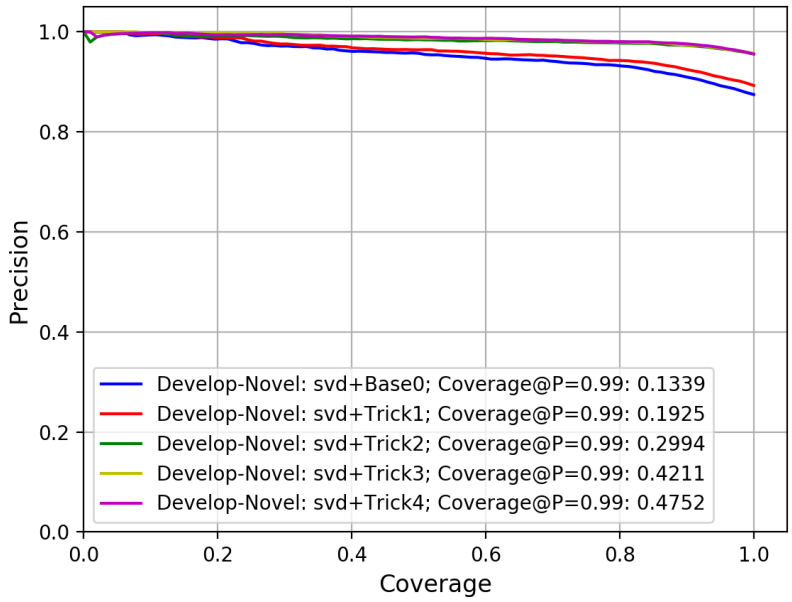

Figure 4: Precision Coverage @P99.

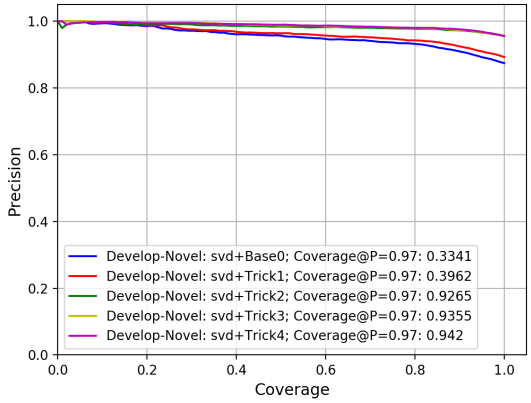

Figure 5: Precision Coverage @P97.

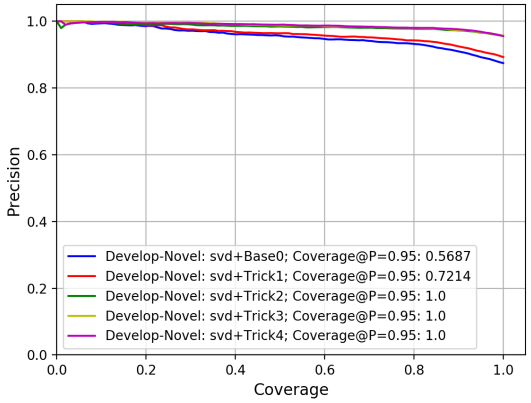

Figure 6: Precision Coverage @P95.

pact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[2] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. PMID: 23964806.

[3] Nawaf Hazim Barnouti, Sinan Sameer Mahmood Aldabbagh, and Wael Esam Matti. Face recognition: A literature review. *International Journal of Applied Information Systems*, 11(4):21–31, Sep 2016.

[4] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *CoRR*, abs/1502.00873, 2015.

[5] Yandong Guo and Lei Zhang. One-shot face recognition by promoting underrepresented classes. Technical report, July 2017.

[6] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. August 2016.

[7] Quan xue Gao, Lei Zhang, and David Zhang. Face recognition using flda with single training image per person. *Applied Mathematics and Computation*, 205(2):726 – 734, 2008. Special Issue on Advanced Intelligent Computing Theory and Methodology in Applied Mathematics and Computation.

[8] Loïc Paulevé, Hervé Jégou, and Laurent Amsaleg. Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31(11):1348–1358, 2010.