Signal Processing: Image Communication 47 (2016) 465–475

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Landmark perturbation-based data augmentation for unconstrained face recognition

Jiang-Jing Lv*, Cheng Cheng, Guo-Dong Tian, Xiang-Dong Zhou, Xi Zhou

Intelligent Multimedia Technique Research Center, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, PR China

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form
31 March 2016
Accepted 31 March 2016
Available online 1 April 2016

Keywords:
Feature representation
Face recognition
Landmark perturbation
Misalignment
Deep convolutional neural networks

ABSTRACT

Face alignment is a key component of face recognition system, and facial landmark points are widely used for face alignment by a number of face recognition systems. However, inaccurate locations of landmark points bring about spatial misalignment which degrades the performance of face recognition systems. In order to alleviate this problem, we propose a simple and efficient data augmentation approach, which uses artificial landmark perturbation to generate a huge number of misaligned face images, to train Deep Convolutional Neural Networks (DCNN) models robust to landmark misalignment. In our experiments, three types of facial landmark-based face alignment methods are applied to train DCNN models on CASIA-WebFace training database. Experimental results on Labeled Faces in the wild database (LFW) and YouTube Faces database (YTF) verify the effectiveness of our approach.

1. Introduction

Automatic face recognition is an important vision task in many practical applications such as identity verification, intelligent visual surveillance and immigration automated clearance system. According to different application scenarios, it can be classified into two different tasks: face verification and face identification. The former aims to determine whether a given pair of face images is from the same person or not, while the latter is to recognize the person from a set of gallery face images and find the most similar one to the probe sample. Many approaches [1–4] have been proposed to improve the face verification performance in unconstrained environments and some of them have exhibited impressive results. For example, Schroff et al. [4] achieved 99.63% face verification accuracy on LFW database [5], which surpasses human accuracy of 97.53%. However, a good verification performance cannot guarantee a good identification performance. Hua et al. [6] concluded that some algorithms have already achieved impressive verification performance on LFW database but get poor performance in identification problem in real environment. In addition, there are still many factors which affect the face recognition performance, such as occlusions, poses, and expressions.

In face recognition systems, face alignment, which tends to warp face images into predefined canonical template, is very critical. Traditionally, facial landmark points are usually used for face alignment and accurate positions of facial landmark points are critical for good recognition performance [7–9]. According to the locations of landmark points, face images can be aligned to predefined canonical template. For instance, Sun et al. [10] aligned face images by using similarity transformation according to the several detected predefined landmark points. Berg et al. [7] incorporated piecewise transformation for face alignment. Taigman et al. [11] introduced a 3D face frontalization method according to the 67 landmark points of 2D face image. If landmark points are accurately located, faces can be well warped and each part of faces from different images will have a good correspondence between each other, which favors feature extracting and feature matching. However, facial landmark detection algorithms are seriously affected by several factors, such as blurring, pose, lighting, expression and occlusion. Fig. 1 shows some examples of misalignment. For a 128 × 128 face image, the alignment error of one landmark point may be up to more than 5 pixels.

Recently, DCNN based feature representation methods, such as FaceNet [4], DeepId [10] and DeepFace [11], have been widely used in face recognition tasks and have shown impressive results in unconstrained environment. Convolutional neural networks (CNN) was first proposed by LeCun et al. in [12] for handwritten code recognition. With large amounts of training data and computation resources such as GPU, since 2012, DCNN have become prevalent and variants of DCNN have been designed in image processing area. For example,

* Corresponding author.
E-mail addresses: lvjiangjing@cigit.ac.cn (J.-J. Lv), chengcheng@cigit.ac.cn (C. Cheng), tianguodong@cigit.ac.cn (G.-D. Tian), zhouxiangdong@cigit.ac.cn (X.-D. Zhou), zhouxi@cigit.ac.cn (X. Zhou).

http://dx.doi.org/10.1016/j.image.2016.03.011
0923-5965/© 2016 Elsevier B.V. All rights reserved.

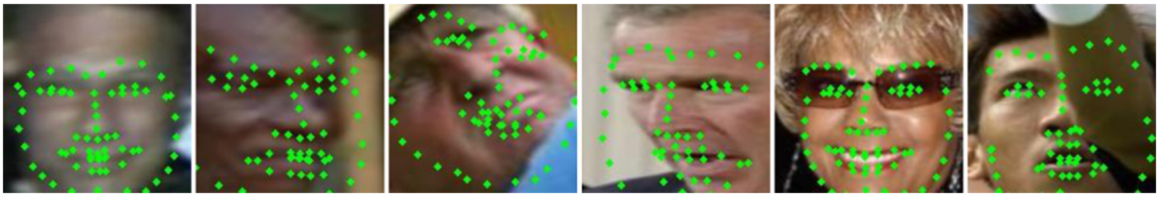


Fig. 1. Examples of misalignment face alignment.

Krizhevsky et al. [13] trained a large and deep convolutional network to classify images and achieved excellent recognition accuracy in ILSVRC-2012 competition [14]. Meanwhile, the architectures of DCNN , such as NIN [15], GoogLeNet [16] and VGG [17], tend to be much deeper and wider, leading to enormous parameters of the network. From Table 1, we can learn that GoogLeNet with 27 layers almost has 16,373K parameters. Thus, training a large DCNN is difficult because it is easy to be over-fitting or even divergency. As shown in Fig. 2, with large network and limited training data, even though the training error is continuously decreasing with increasement of epoches, the test error is increasing after several epoches. A large number of strategies have been proposed to address this problem. On the one hand, different regularization methods have been adopted to DCNN training, such as Dropout [18], Maxout [19] and DropConnect [20]. On the other hand, collecting more training data can essentially deal with this problem. Better performance can be achieved with more training data, however, it is difficult and expensive to collect a large number of labeled data. Therefore, data augmentation strategies, such as flipping [13], cropping [13,21], color casting [22] and blurring [23], have been proposed, which artificially generate large number of visual training data, and experimental results show that data augmentation can help the trained model get a strong generalization ability to unseen but similar patterns in the training data.

Inspired by data augmentation methods, we propose a simple and efficient landmark perturbation-based data augmentation method to alleviate the problem of misalignment. It automatically perturbs the landmark positions to generate a huge number of misaligned face images to train DCNN model, examples of landmark perturbation are shown in Fig. 3. There are some prior works similar to our work. In [24], the authors used several data augmentation methods to generate more training data, including flipping, shifting, rotation, scaling and cropping. In [25], data augmentation were used in augmentation of landmark points. Besides flipping and rotation, the authors also added Gaussian noise to the raw landmark points to generate more examples. Although related, our approach is different from previous ones in several ways: we can automatically generate different kinds of images (e.g., translation, rotation, scaling and shear) by using landmark perturbation-based data augmentation without complex composing of different data augmentation methods; different from [25] which just augments landmark points and extracts geometry feature of landmark points for facial expression recognition, we use the perturbed landmark points for face alignment and aim to generate more face images with misalignment for DCNN training; and experimental results on LFW [5] and YTF [26] show that the DCNN models trained by our method are robust to misalignment and significantly improve the face recognition rates.

The rest of this paper is organized as follows. Section 2 reviews some related previous works. Section 3 presents our data augmentation approach. Experimental setup and results are presented in Section 4. Section 5 offers our conclusions.

Table 1
The architecture of the GoogLeNet.

Name	Type	Filter size/ stride	Output size	Depth	#Params
Conv11	Convolution	7 × 7/2	64 × 64 × 64	1	2.7K
Pool1	Max pooling	3 × 3/2	32 × 32 × 64	0	
Conv21	Convolution	3 × 3/1	32 × 32 × 192	2	112K
Pool2	Max pooling	3 × 3/2	16 × 16 × 192	0	
Inception3a	Inception		16 × 16 × 256	2	159K
Inception3b	Inception		16 × 16 × 480	2	480K
Pool3	Max pooling	3 × 3/2	8 × 8 × 480	0	
Inception4a	Inception		8 × 8 × 512	2	364K
Inception4b	Inception		8 × 8 × 512	2	437K
Inception4c	Inception		8 × 8 × 512	2	463K
Inception4d	Inception		8 × 8 × 528	2	580K
Inception4e	Inception		8 × 8 × 832	2	840K
Pool4	max pooling	3 × 3/2	4 × 4 × 832	0	
Inception5a	Inception		4 × 4 × 832	2	1072K
Inception5b	Inception		4 × 4 × 1024	2	1388K
Pool5	Avg pooling	5 × 5/1	1 × 1 × 1024	0	
Linear1	Fully connection		1 × 1 × 10,575	1	10,575K
Cost	Softmax		1 × 1 × 10,575	0	
Total				22	16,373K

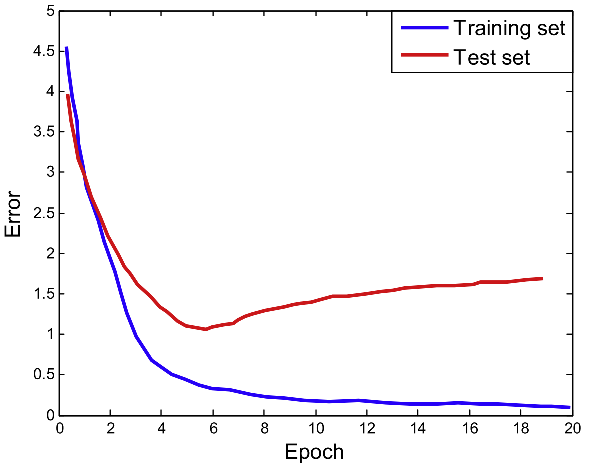


Fig. 2. The error curves of training set and test set with increasing of epoches.

2. Related work

The curse of misalignment in face recognition was first proposed by Shan et al. in [27], which systematically evaluated Fisherface's sensitivity to misalignment problem by perturbing the eye coordinates and revealed that imprecise localization of the facial landmarks would abruptly degenerate the Fisherface system. Additionally, misalignment, which enlarges the within-class scatter and reduces the between-class scatter to some degree, increases the difficulties of face recognition. Many face recognition methods suffer from misalignment problem. Sparse representation-based classification (SRC) [28], which seeks a sparse linear representation of the probe images over the training images, is also sensitive to misalignment. Feature encoding methods, such as Fisher vector [29] and Hierarchical Gaussianization Vector [30], also need well-aligned images as the input.

In order to overcome the curse of misalignment problem, large number of methods have been proposed. Shan et al. [27] proposed

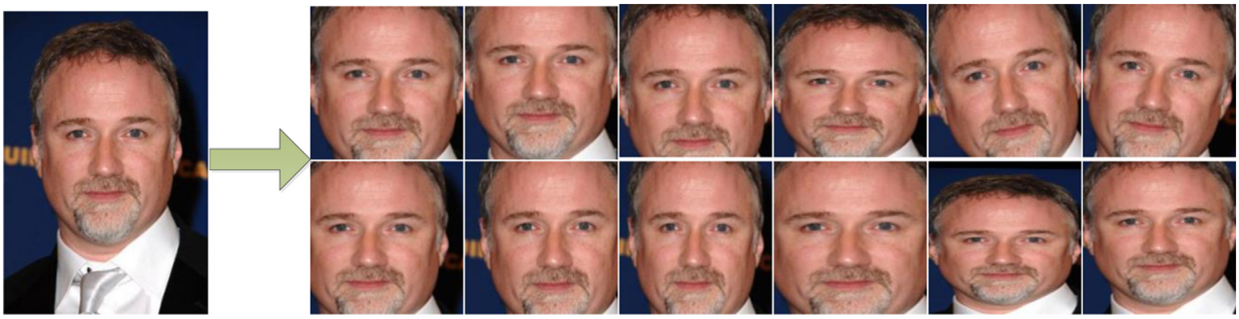


Fig. 3. Examples of landmark perturbation for face alignment.

an E-Fisherface misalignment learning method, which reinforces the recognizer to model the misalignment variations. Yang et al. [31] proposed a subspace learning method by solving a constrained l_1 norm optimization problem, through which the underlying spatial misalignment parameters of face image were learned. Sparse representation techniques, which have been widely studied in the past several years, were also applied to face recognition with misalignment. Robust alignment by sparse representation (RASR) [28], which aligned test images with training images, obtained impressive results. But RASR searches the best aligned test image via a subject-to-subject optimization which is time-consuming for large-scale and real-time face recognition systems. Yang et al. [31] presented an efficient misalignment-robust representation (MRR) for real-time face recognition systems via correspondence-based representation and a coarse-to-fine optimization. Though MRR has achieved impressive results with low computation cost, it is not robust enough to deal with images with occlusion and illumination variations. Derived from MRR, Tai et al. [32] presented a novel method called structure constraint coding (SCC), which uses the unclear norm as structure constraint on the error matrix to keep structural information of images. Experimental results demonstrate that SCC handles with face misalignment coupled with noise better than that of MRR.

While most previous works focus on frontal images in constrained environment with small dataset, it is hard to deal with images which have large pose, occlusion and illumination variance. In view of this situation, a number of works [7,9–11] paid attention to aligning face images to a fixed canonical template. Most of them use landmark detector algorithm to get locations of predefined landmark points in the images, then geometry transformations are used to align face images to a predefined canonical template. However, these methods strongly depend on accurate localization of the positions of landmark. As shown in Fig. 1, the positions of landmark sometimes cannot be accurately located.

In contrast to previous works which aim to get images precisely aligned, we aim to learn an invariant feature which is robust to misalignment. By taking advantage of DCNN's powerful representation ability and inspired from previous data augmentation methods [13,22–24,33,25], we propose a landmark perturbation-based data augmentation method to generate a large set of misaligned images to train DCNN model through which the feature extracted is robust to misalignment.

3. Landmark perturbation

In modern face recognition systems, the face recognition pipeline usually consists of face detection, face alignment, feature representation and feature matching. Face alignment is a critical image preprocess step in the process, which affects the performance of face recognition to some degree. During face alignment, the locations of landmark points are detected and all faces are aligned to the predefined canonical template. With landmark points accurately located, faces can be well warped and each part of faces has a good correspondence between each other, which contributes to the following feature extracting and feature matching.

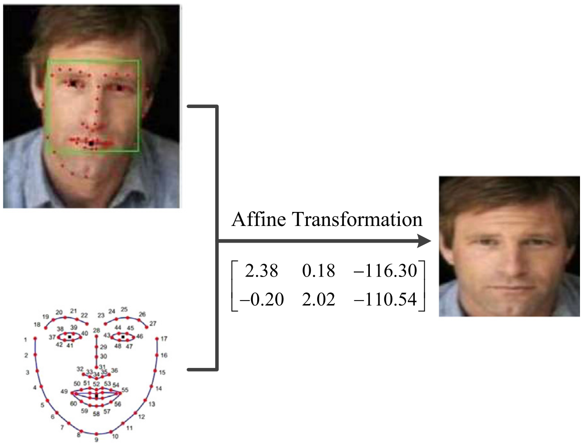


Fig. 4. Process of affine transformation.

Traditionally, affine transformation is widely used for face alignment and the centers of the eyes and mouth are the three fiducial points which are used for parameters calculation. The process of affine transformation is shown in Fig. 4. The general affine transformation function is defined as:

$$x_{dest} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} * x_{source} + t \tag{1}$$

where x_{source} is a facial landmark in the input face image, x_{dest} is a well-defined facial landmark on the standard face, $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ and t are transformation parameters. By solving Eq. (1), we get the transformation parameters. Then, the input face image can be mapped to the fixed canonical template.

If the positions of facial landmark points are accurately located, the face image can be well aligned. But in practice, as shown in Fig. 1, due to occlusion, blurring, pose and some other factors, the misalignment is almost unavoidable. In order to train a DCNN model which is robust to misalignment, images with various transformations (e.g., translation, rotation, shearing and scaling) are needed to train the DCNN model. These transformations can be obtained by changing the values of the affine parameters.

Suppose that I is the original face image and I_0 is the warped face image by affine transformation:

$$I_0 = I \odot A \tag{2}$$

where A is a 3×3 affine transformation matrix. According to the characters of affine transformation, A can be decomposed into the composition of rotation, shearing, scaling and translation transformations:

$$A = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & shx & 0 \\ shy & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} sx & 0 & 0 \\ 0 & sy & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 0 & dx \\ 0 & 1 & dy \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

where θ is the rotation angle, shx and shy are the shearing ranges, sx and sy are the scaling parameters, dx and dy control the translation over x-axis and y-axis. So we can adjust these parameters to get any kind of misaligned images. But the decomposition of A is complex by using QR decomposition and multiple matrix multiplication are required to generate each misaligned image. Both of the two steps are time-consuming.

Meanwhile, we find that the misalignment of the landmark points is equivalent to the variation of the affine parameters. So in practice, we derive multiple misaligned virtual face samples by perturbing each image's landmark points' coordinates in a certain range and use the new coordinates to calculate the matrix A. The landmark perturbation strategy can be defined as:

$$P_i^* = P_i + r \tag{4}$$

where P_i is the localized position and P_i^* is the new position with perturbation by adding random perturbation range r.

The random perturbation range can be Gaussian or Uniform distributed:

$$r \sim N(\mu, \Sigma) \tag{5}$$

$$r \sim U(-a, a) \tag{6}$$

For both distributions, the mean is set as 0 and covariance matrix is a diagonal matrix sharing the same variance σ^2.

The processes of landmark perturbation-based data augmentation method is as follows.

Firstly, face regions are detected by a Viola and Jones [34] based face detector and resized to fixed size of 160 × 160. Secondly, the SDM algorithm [35] is adopted to detect 68 facial landmark points $\{(x_1, y_1), \ldots, (x_{68}, y_{68})\}$. Thirdly, the locations of landmark are randomly perturbed within a certain range to generate a series of sets of landmark points for face alignment. Lastly, face images are aligned by these series of sets of landmarks respectively. The pipeline of landmark perturbation-based data augmentation method is shown in Fig. 5.

The landmark perturbation-based data augmentation can not only be used in affine transformation, but also be suitable to other transformation methods, such as linear conformal transformation and piecewise linear transformation which are also widely used for face alignment. The comparison and details of these transformations are as follows.

Linear conformal transformation is similar with the affine transformation, which only use the centers of the eyes for parameter calculation. The general linear conformal transformation can be defined as:

$$x_{dest} = \begin{bmatrix} a & -b \\ pb & pa \end{bmatrix} * x_{source} + t \tag{7}$$

where p is a parity parameter which set to be 1 or -1.

Piecewise linear transformation makes affine transformation in each piece, which can be defined as:

$$x_{dest}^i = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} * x_{source}^i + t_i, \quad i = 1, \ldots, N \tag{8}$$

where N is the piece number, x^i is the point in region i. Fig. 6 shows the corresponding Delaunay triangulation and the process of piecewise linear transformation. We divide face image into 87 triangles and each triangle is aligned by affine transformation. Each piece of face between images is corresponded and images are transformed to frontal which reduces the effects of pose and expression variance.

The examples generated by the three face alignment methods are shown in Fig. 7. Landmark perturbation-based data augmentation method artificially generates a huge number of misaligned face images, including stretched, distorted, rotated, cropped, which makes the trained model robust to these factors.

4. Experiments

4.1. Feature representation

In the past few years, DCNN have achieved outstanding performance on image classification and object detection tasks. There are many tricks for DCNN training with very deep architecture

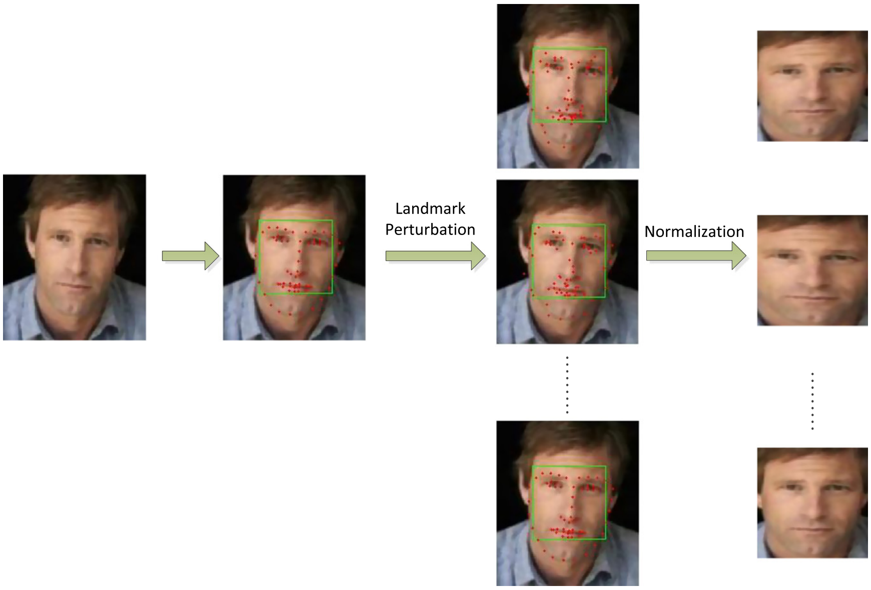


Fig. 5. The pipeline of landmark perturbation for face alignment.

[17,16,15], non-saturated activation function [36] and dropout [37]. Inspired by the outstanding work of Szegedy et al. [16], which introduced GoogLeNet architecture and won the ImageNet 2014 competition, we also use the GoogLeNet architecture to train the feature extraction model in our experiments. GoogLeNet is a very deep architecture, which contains 22 layers with 16,373K parameters and 5 pooling layers. Furthermore, inspired by a neuroscience model of the primate visual cortex, it introduces inception layer to its architecture, which has filters of different sizes (e.g., 1 × 1, 3 × 3, 5 × 5) to find the optimal local construction.

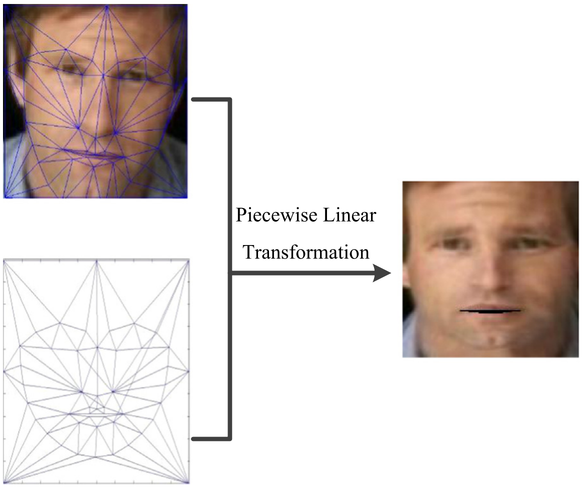


Fig. 6. Process of piecewise linear transformation.

The configuration of GoogLeNet is shown in Table 1. It uses ReLU [38] activation function $\max(0, x)$ after every convolution and fully connected layer. Max-pooling layer takes the max over each 3×3 spatial neighborhoods with a stride of 2. Softmax function is used as loss function which computes the probability of k-th output assigned to the k-th class by $p_k = \exp(o_k)/\Sigma_h \exp(o_h)$. The goal is to maximize the probability of the correct class prediction by standard back-propagation [39]. The special property of deep learning feature is low-dimension and sparse with strong discrimination. Therefore, we use GoogleNet to extract image features throughout our experiments.

4.2. Experiment setup

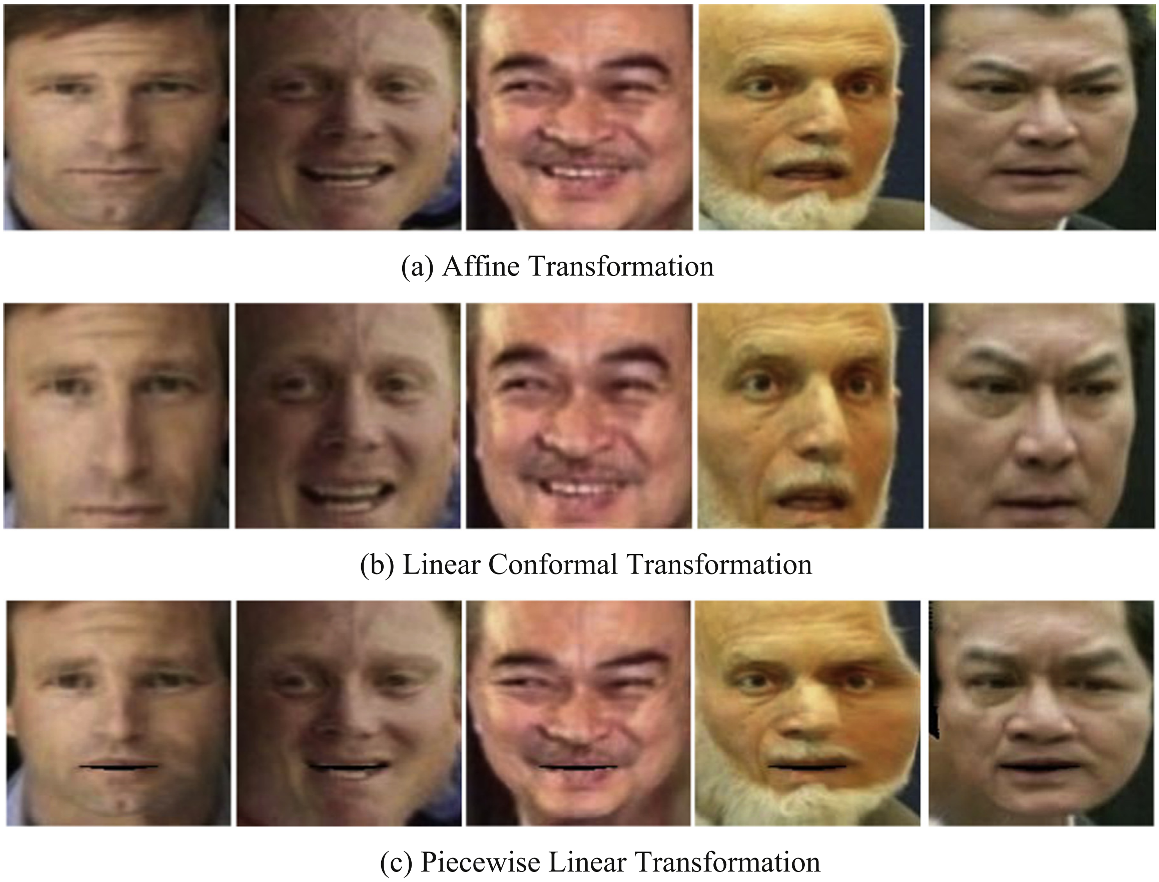
We adopt public available deep learning framework Caffe [40] to train our GoogLeNet models.

During training stage, we use the poly rule to dynamically adjust the learning rate.

$$lr\_rate = base\_lr_{*}\left(1 - \frac{iter}{iter\_num}\right)^{gamma} \tag{9}$$

where $base\_lr$ is the start learning rate, $iter$ is the current iteration number and $iter\_num$ is the max iteration number. In our experiments, we start with a learning rate of 0.01 with poly rule to finalize the model. The max iteration number is set to 500,000, and 700,000 for data augmentation. The $gamma$ is set to 0.01. On the other hand, we use small mini-batches, which set to be 128, to improve the convergence during Stochastic Gradient Descent (SGD) [41] backward propagation.

The GoogLeNet is trained on CASIA-WebFace dataset [42], which contains 10,575 subjects of 494,414 images. The SDM



(a) Affine Transformation

(b) Linear Conformal Transformation

(c) Piecewise Linear Transformation

Fig. 7. Examples of different face alignment methods.

algorithm [35] is used to detect 68 facial landmark points for each face image in CASIA-WebFace dataset. According to the detected landmark points, landmark perturbation and image alignment are conducted. In our experiments, images were aligned to 128 × 128 and Uniform distribution with $\sigma^2 = 5$ was use for perturbation for simplicity. Different perturbation ranges and distributions will be discussed in Section 4.6. For the similarity of some virtual images, we generated almost 200 virtual samples for each training image, which almost cover all the situations. In order to compare the contribution of landmark perturbation to different alignment methods, the following DCNN models are trained respectively:

- **A:** Affine transformation.
- **A_P:** Affine transformation + Landmark Perturbation.
- **B:** Linear conformal transformation.
- **B_P:** Linear conformal transformation + Landmark perturbation.
- **C:** Piecewise linear transformation.
- **C_P:** Piecewise linear transformation + Landmark Perturbation.

In the phase of face verification or identification, the Joint Bayesian approach [43] which has been successfully applied to previous face recognition systems [10,44], was used for feature matching throughout our experiments.

In order to verify the effectiveness of our method, we evaluated the proposed method on LFW database [5] in both verification and identification scenarios and YTF [26] database with verification condition.

4.3. Experiment results on LFW

4.3.1. Experiment results under standard protocol of LFW

We followed the unrestricted setting of LFW and measured the performance of each model using the 10-fold cross-validation. In particular, we trained the Joint Bayesian model on 9 splits, and tested it on the remaining split. Each split contains 600 image pairs which were predefined by LFW. Mean accuracy and standard error were reported.

Firstly, we used the SDM algorithm [35] to detect 68 facial landmark points of each LFW image. Then each image in LFW was aligned by the alignment methods mentioned above. Lastly, we tested the performance of different DCNN models, which were trained by different alignment methods using landmark perturbation or not. The results are listed in Table 2 and ROC curves are shown in Fig. 8, where Fusion is the performance by cascading the features of above three transformation methods. When comparing results between LP and No-LP, where landmark perturbation-based data augmentation method is employed in the training phrase or not, we find that the accuracy rates are greatly improved by LP. Especially for linear transformation, it reduces the error up to 29%.

When compared with previous results on LFW, we achieved 99.28% by fusing the models trained by LP, which is much better than others except FaceNet [4]. Because FaceNet was trained on the enormous database containing 200 million photos of 8 million people which is a thousand times larger than our training database. Furthermore, it adopted triplet loss function which has a strong discriminative ability. Therefore, if more training data and triplet loss function were added to our model training, it is quite possible to obtain much better result.

Table 2
The performance of different transformation methods under standard protocol.

Train	Test	Accuracy (%)	
		No-LP	LP
Affine	Affine	98.63 ± 0.51	98.97 ± 0.47
Linear	Linear	98.13 ± 0.47	98.67 ± 0.56
Piecewise	Piecewise	97.65 ± 0.76	98.03 ± 0.60
Fusion		98.98 ± 0.39	99.28 ± 0.41
DeepFace [45]		98.00	
DeepID2+ [46]		98.70	
DR+Joint Bayse [42]		97.73	
FaceNet [4]		99.63	

4.3.2. Experiment results under BLUFR protocol of LFW

We also evaluated our proposed method in face identification scenario, following the evaluation protocol of [44] which contains both verification and open-set identification scenarios, with a focus at low false accept rates (FARs). It adopted 10 random trials of training and testing. For each of the 10 trials, the whole LFW database was randomly divided into a training set and a testing set. The training set of each trial includes 1500 subject, 3524 images on average, while the testing set contains the remaining 4249 subjects, 9708 images on average. The average accuracy of 10 trails were reported. Using the toolkit provided by [44], we tested the performance of our landmark perturbation-based data augmentation method in the open-set identification scenario. The results of open-set identification at rank 1 with FAR=1% are listed in Table 3 and the ROC curves with rank 1 are shown in Fig. 9. From Table 3 and Fig. 9, we note that the proposed landmark perturbation-based data augmentation method also improves the face identification rates significantly. Compared with previous result in [42], we achieved 63.73% by fusing the models trained by LP reducing the error by 49%.

4.4. Experiment results on YTF

We further tested the generalization capability of our proposed method on YouTube Faces (YTF) database [26], which contains 3425 videos of 1595 different people collected from YouTube site. We followed the standard evaluation protocol and tested on the predefined 5000 video pairs. The 5000 video pairs are divided into 10 splits, each of them containing 250 pairs from the same person and 250 pairs of different persons. Similar to LFW, we trained the Joint Bayesian model on 9 splits, and tested it on the remaining split.

Table 4 and Fig. 10 show the face verification performance on YTF database. We note that landmark perturbation-based data augmentation method improves the performance significantly, and the best performance is achieved by fusing the features of the models trained by landmark perturbation. Compared to the previous results reported on this test set, 91.40% of verification rate (VR) in [11], 92.24% of VR in [42], and 93.20% of VR in [46], the proposed approach achieves 94.04% of VR, demonstrating significant improvement and advantage. Even though FaceNet [4] achieved the best accuracy, it benefitted from its triplet loss function and larger training set.

4.5. Comparison with other data augmentation methods

We compare the performance with widely used data augmentation methods, such as flipping [13], patches (clipping) [13,33,21], color casting [22], blurring [23]. Examples are shown in Fig. 11.

In order to compare different data augmentation methods, the following GoogLeNet models were trained respectively. As contrast, blurring, noise and color casting are simple and common used data augmentation methods, we mixed these augmentation methods and trained only one model for convenience:

- **Baseline**: No data augmentation.
- **A**: Flipping.
- **B**: Contrast + Blurring + Noise + Color casting.

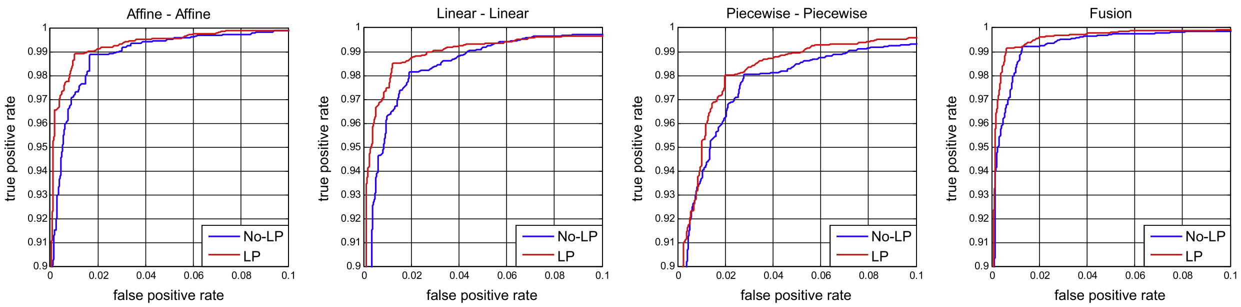


Fig. 8. ROC curves of different transformation methods under standard protocol.

Table 3
The performance of different transformation methods under BLUFR protocol. The reported numbers are the mean detection and identification rates (%) at rank 1 with FAR=1%.

Train	Test	Accuracy (%)	
		No-LP	LP
Affine	Affine	52.21	57.90
Linear	Linear	42.22	53.47
Piecewise	Piecewise	37.07	42.87
Fusion		53.90	63.73
DR+Joint Bayse [42]		28.90%	

- **C**: Patches.
- **D**: Landmark perturbation.

Table 5 shows the face recognition rates on LFW and YTF databases. Among all these data augmentation methods, patches and landmark perturbation perform much better than other methods, while landmark perturbation achieves highest rates on verification task. We further explore the performance of fusing different data augmentation models. In this case, we just concatenate the features extracted by different models. We find that with incorporating more models the performance can be significantly improved. Even the fused models have achieved high recognition rates, adding landmark perturbation model can still improve the face recognition rates. This test shows that landmark perturbation is practical and efficient, and can be used to improve the face recognition performance.

4.6. Discussion

In this section, we provide more comprehensive analysis of the proposed landmark perturbation-based data augmentation method.

First, we tested the relationship between recognition rates and iteration number. Affine transformation was used to align face image in this experiment and experimental results are shown in Fig. 12. We find that LP improves convergence speed of neural network effectively and with increasing of iteration number the performance stably increased. While No-LP is relatively slower and after iteration number reached 500,000, the recognition rates are oscillated with little improvement. For the same recognition rate, LP needs less training iterations than No-LP. These comparisons show that LP provides faster convergence at early stage and better performance can be achieved with more iteration.

Second, the performance with various distributions and variances were tested. Empirically, for inaccurate locations of landmark points, they are usually densely distributed around the ground truth. If the generated visual images are much closer to the practical situation, the trained DCNN model may have better generalization ability. Therefore Gaussian distribution is employed to imitate the distribution, while Uniform distribution is used for comparison. Furthermore, variances are used to estimate the alignment error. According to (5) and (6), the DCNN models were trained by affine transformation combined with landmark perturbation respectively. We compare $\sigma^2 = \{0, 5, 10, 20\}$, where 0 means without perturbation. The results are plotted in Fig. 13. As shown, when $\sigma^2 = 5$, Gaussian distribution gets the best performance and the performance decrease when $\sigma^2 > 5$. While Uniform distribution achieves the best verification rate at $\sigma^2 = 10$ and identification at $\sigma^2 = 5$. In general, the differences between Gaussian distribution and Uniform distribution are not obvious in face verification, which is less than 0.3% in verification rate. While Gaussian distribution achieves better performance than Uniform distribution in identification, especially $\sigma^2 = 5$ Gaussian distribution gets over 2% relative improvement. For higher variances, the performance of both distributions are declined. We argue that for larger variance the face images are seriously distorted which affects discriminative model training.

Lastly, we tested the face recognition performance under extreme situation where the detected face regions were directly employed without using any alignment. First, the detected face regions were resized to the fixed size. Second, features corresponding to different pre-trained models were extracted respectively. Lastly, the verification and identification rates were obtained according to the above protocols. Experimental results on LFW database are listed in Tables 6 and 7 and the ROC curves are shown in Figs. 14 and 15. We note that landmark perturbation-based data augmentation method can dramatically improve the performance, especially affine transformation model which reduces the error by 53.5% under standard protocol and 15.3% under BLUFR protocol. While piecewise transformation model which was trained with frontal images gets little improvement for the detected face images having large pose and expression variations.

5. Conclusion

This paper presents a landmark perturbation-based data augmentation approach for DCNN training. It is easy to implement by randomly perturbing the locations of landmark points during face alignment. The artificial visual images including different kinds of transformed images which greatly enlarge the training set make the DCNN robust to factors (e.g., stretched, distorted, rotated and cropped) and provide faster convergence for DCNN training at early stage. Experimental results in unconstrained face recognition demonstrate that the features extracted by DCNN models trained with LP can significantly improve the face recognition performance both in verification and identification scenarios. In particular, under BLUFR protocol of LFW, we attained an identification rate of 63.73% which is significantly higher than those reported in the literatures.

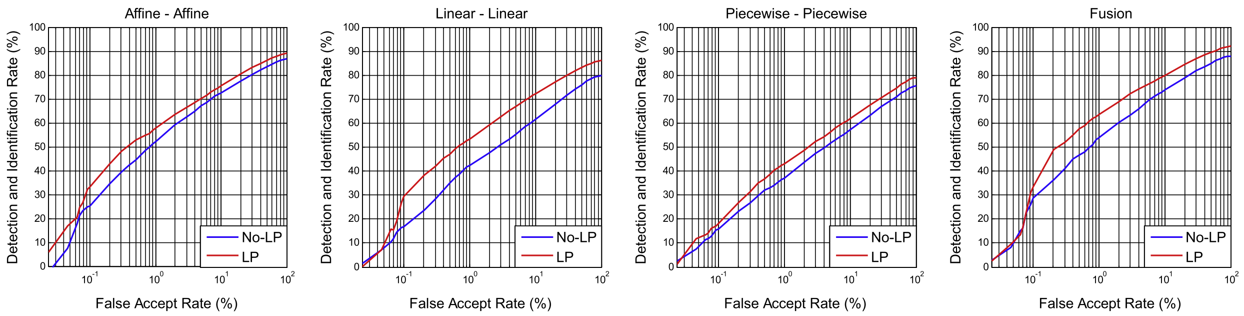


Fig. 9. Identification ROC curves of different transformation methods under BLUFR protocol.

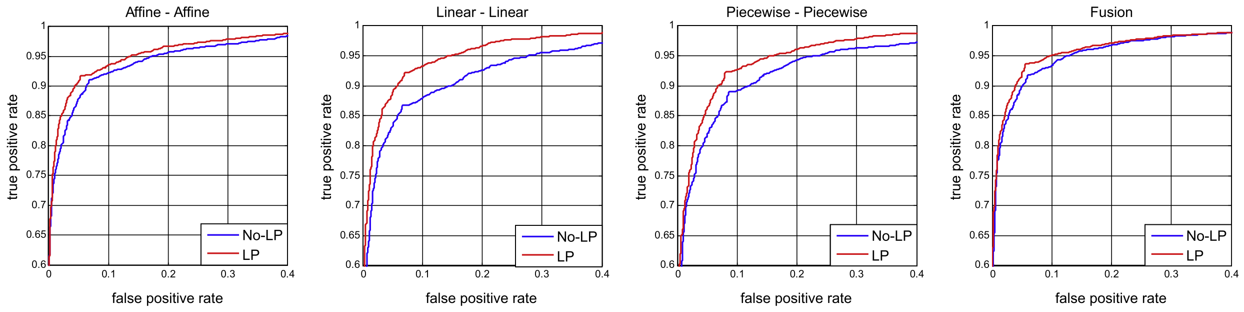


Fig. 10. ROC curves of different transformation methods on YTF.

Table 4
The performance of different transformation methods on YTF.

Train	Test	Accuracy (%)	
		No-LP	LP
Affine	Affine	92.18 ± 0.97	93.24 ± 0.99
Linear	Linear	90.06 ± 0.38	92.62 ± 0.98
Piecewise	Piecewise	90.22 ± 0.63	92.14 ± 1.06
Fusion		93.00 ± 1.11	94.04 ± 1.49
DeepFace [11]		91.40	
DR+Joint Bayse [42]		92.24	
FaceNet [4]		95.12	
DeepID2+ [46]		93.20	

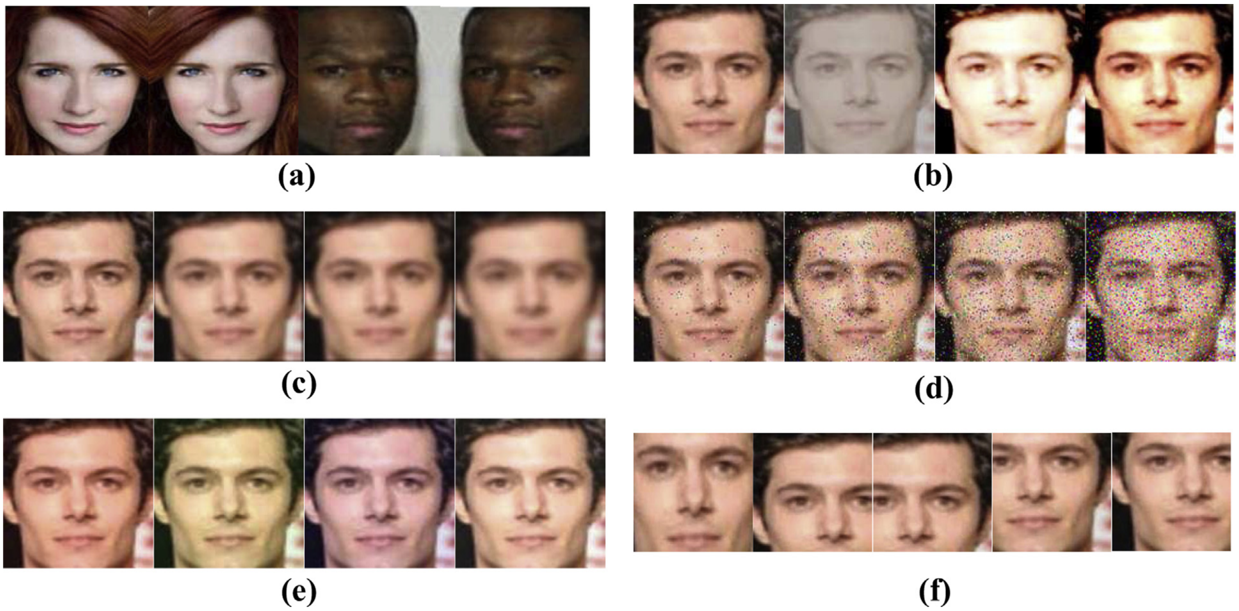

(a)
(b)
(c)
(d)
(e)
(f)

Fig. 11. Examples of a different data augmentation. (a) Flipping, (b) Contrast, (c) Blurring, (d) Noise, (e) Color casting, (d) Patches.

Table 5
Recognition rates (%) with various data augmentation methods.

Methods	LFW		YTF
	Standard	BLUFR	Accuracy
Baseline	98.22 ± 0.62	47.17	92.08 ± 1.29
A	98.63 ± 0.51	52.21	92.18 ± 0.97
B	98.77 ± 0.55	55.42	92.66 ± 1.41
C	99.07 ± 0.45	67.81	93.46 ± 0.99
D	99.28 ± 0.41	63.73	94.04 ± 1.49
Fusion {**A,B**}	99.00 ± 0.47	58.69	93.46 ± 1.27
Fusion {**A,B,C**}	99.18 ± 0.43	70.08	93.90 ± 1.32
Fusion {**A,B,C,D**}	99.35 ± 0.36	71.02	94.28 ± 1.41

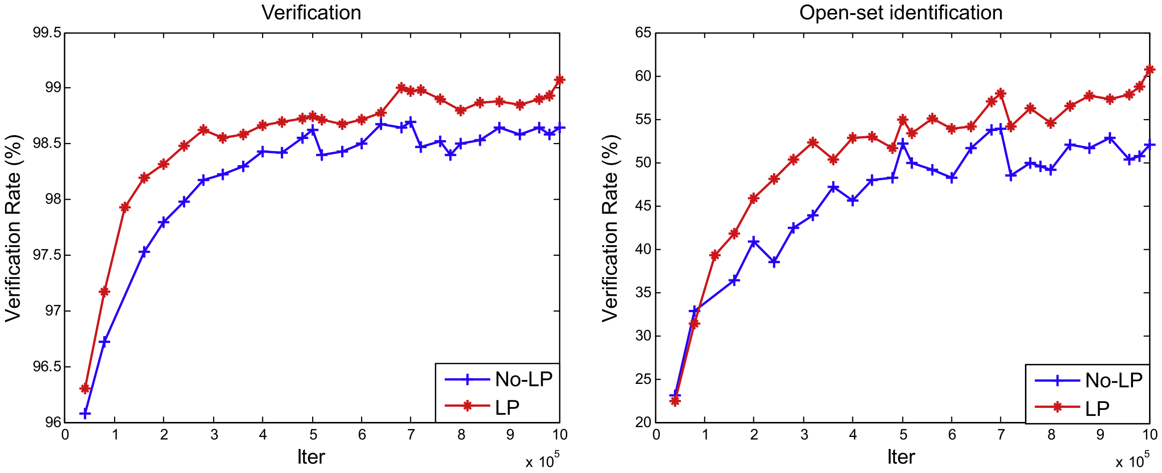


Fig. 12. Face recognition rates on LFW database with increasing of iteration number.

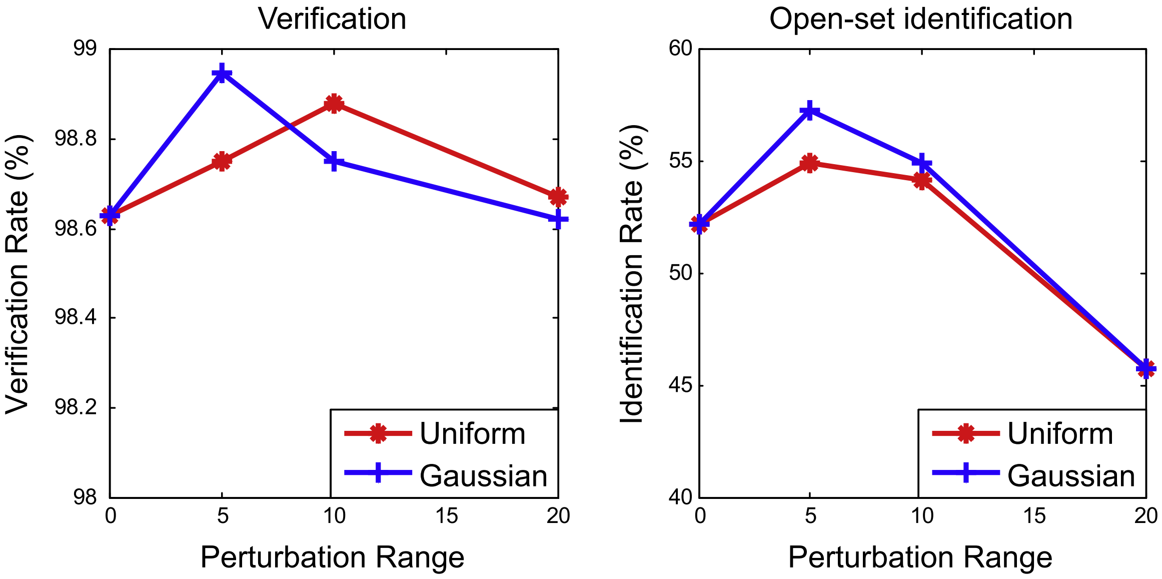


Fig. 13. Curves of recognition rate on LFW with various distributions and variances.

Table 6
The performance of using detected face images with different transformation methods under standard protocol.

Method	Accuracy (%)	
	No-LP	LP
Affine	92.37 ± 1.01	96.45 ± 0.81
Linear	96.33 ± 0.39	97.87 ± 0.51
Piecewise	88.97 ± 1.20	90.13 ± 1.14
Fusion	96.62 ± 0.75	97.93 ± 0.60

Table 7
The performance of using detected face images with different transformation methods under BLUFR protocol. The reported numbers are the mean detection and identification rates (%) at rank 1 with FAR = 1%.

Method	Accuracy (%)	
	Baseline	LP
Affine	9.30	23.21
Linear	23.79	39.42
Piecewise	4.19	5.41
Fusion	26.10	39.23

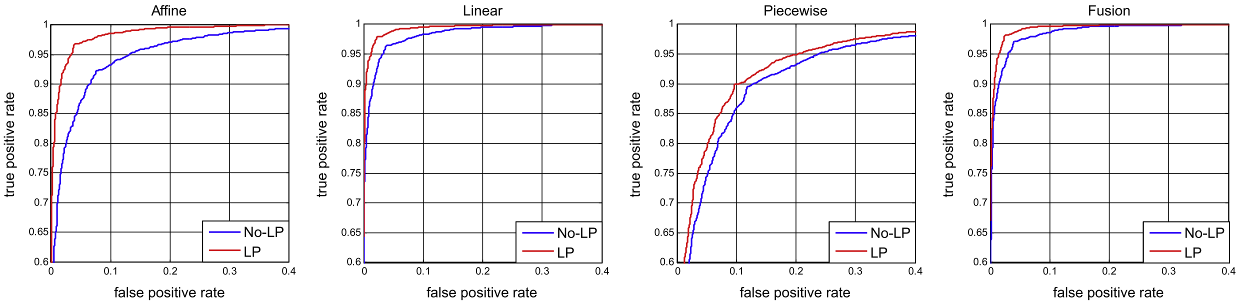


Fig. 14. ROC curves of using detected face images with different transformation methods under standard protocol.

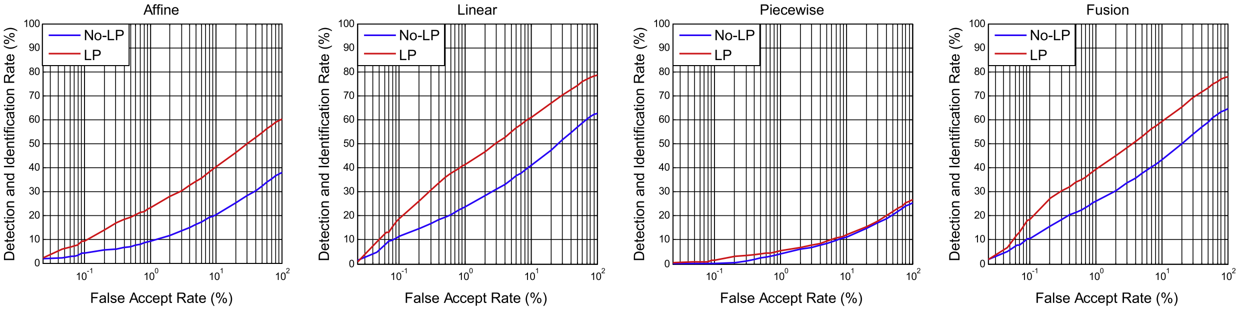


Fig. 15. Identification ROC curves of using detected face images with different transformation methods under BLUFR protocol.

Furthermore, LP can be combined with others data augmentation methods to improve the face recognition performance.

Acknowledgments

This work was partially supported by the National Natural Science Foundations of China (Grant nos. 61472386 and 61502444). The authors would like to thank You-ji Feng for valuable discussions.

Appendix A. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.image.2016.03.011.

References

[1] S.U. Hussain, T. Napoléon, F. Jurie, Face recognition using local quantized patterns, in: British Machine Vision Conference, 2012, 11 pp.

[2] Z. Lei, D. Yi, S.Z. Li, Local gradient order pattern for face representation and recognition, in: International Conference on Pattern Recognition, 2014, pp. 387–392.

[3] J. Ylioinas, A. Hadid, J. Kannala, M. Pietikainen, An in-depth examination of local binary descriptors in unconstrained face recognition, in: International Conference on Pattern Recognition, 2014, pp. 4471–4476.

[4] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[5] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report 07–49, University of Massachusetts, Amherst, 2007.

[6] G. Hua, M.-H. Yang, E. Learned-Miller, Y. Ma, M. Turk, D.J. Kriegman, T. S. Huang, Introduction to the special section on real-world face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (10) (2011) 1921–1924.

[7] T. Berg, P.N. Belhumeur, Tom-vs-pete classifiers and identity-preserving alignment for face verification, in: British Machine Vision Conference, vol. 2, 2012, p. 7.

[8] H.K. Ekenel, R. Stiefelhagen, Why is facial occlusion a challenging problem? in: Advances in Biometrics, Springer, Berlin Heidelberg, 2009, pp. 299–308.

[9] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, Int. J. Comput. Vis. 107 (2) (2014) 177–190.

[10] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891–1898.

[11] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.

[12] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.

[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248–255.

[15] M. Lin, Q. Chen, S. Yan, Network in network, arXiv preprint arXiv:1312.4400.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[17] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, arXiv preprint arXiv:1405.3531.

[18] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, arXiv preprint arXiv:1207.0580.

[19] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, arXiv preprint arXiv:1302.4389.

[20] L. Wan, M. Zeiler, S. Zhang, Y.L. Cun, R. Fergus, Regularization of neural networks using dropconnect, in: International Conference on Machine Learning, 2013, pp. 1058–1066.

[21] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, Proc. Br. Mach. Vis. 1 (3) (2015) 6.

[22] R. Wu, S. Yan, Y. Shan, Q. Dang, G. Sun, Deep image: scaling up image recognition, arXiv preprint arXiv:1501.02876.

[23] T. Ahonen, E. Rahtu, V. Ojansivu, J. Heikkila, Recognition of blurred faces using local phase quantization, in: International Conference on Pattern Recognition, 2008, pp. 1–4.

[24] T. Devries, K. Biswaranjan, G.W. Taylor, Multi-task learning of facial landmarks and expression, in: Canadian Conference on Computer and Robot Vision, Montreal, QC, 2014, pp. 98–103.

[25] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, J. Kim, Deep temporal appearance-geometry network for facial expression recognition, arXiv preprint arXiv:1503.01532.

[26] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 529–534.

[27] S. Shan, Y. Chang, W. Gao, B. Cao, P. Yang, Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution, in: IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, South Korea, 2004, pp. 314–320.

[28] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.

[29] K. Simonyan, O.M. Parkhi, A. Vedaldi, A. Zisserman, Fisher vector faces in the wild, in: British Machine Vision Conference, 2013.

[30] X. Zhou, N. Cui, Z. Li, F. Liang, T.S. Huang, Hierarchical Gaussianization for image classification, in: IEEE International Conference on Computer Vision, 2009, pp. 1971–1977.

[31] M. Yang, L. Zhang, D. Zhang, Efficient misalignment-robust representation for real-time face recognition, in: European Conference on Computer Vision, Springer, Berlin Heidelberg, 2012, pp. 850–863.

[32] Y. Tai, J. Qian, J. Yang, Z. Jin, Face recognition with image misalignment via structure constraint coding, in: Asian Conference on Computer Vision, Springer, International Publishing, 2014, pp. 558–573.

[33] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1805–1812.

[34] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, pp. I–511.

[35] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.

[36] X. Bing, W. Naiyan, C. Tianqi, L. Mu, Empirical evaluation of rectified activations in convolution network, arXiv preprint arXiv:1503.03832.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, in: The Journal of Machine Learning Research, vol. 15, 2014, pp. 1929–1958.

[38] G. Dahl, T. Sainath, G. Hinton, Improving deep neural networks for lvcsr using rectified linear units and dropout, in: International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8609–8613.

[39] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by backpropagating errors, Cogn. Model. 5 (1988).

[40] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the ACM International Conference on Multimedia, ACM, New York, NY, USA, 2014, pp. 675–678.

[41] D.R. Wilson, T.R. Martinez, The general inefficiency of batch training for gradient descent learning, Neural Netw. 16 (10) (2003) 1429–1451.

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

[43] D. Chen, X. Cao, F. Wen, J. Sun, Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3025–3032.

[44] S. Liao, Z. Lei, D. Yi, S.Z. Li, A benchmark study of large-scale unconstrained face recognition, in: IEEE International Joint Conference on Biometrics, 2014, pp. 1–8.

[45] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Web-scale training for face identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[46] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.