Neurocomputing 230 (2017) 184–196

Contents lists available at ScienceDirect

Neurocomputing

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

Data augmentation for face recognition

Jiang-Jing Lv[a,b], Xiao-Hu Shao[b], Jia-Shui Huang[b], Xiang-Dong Zhou[b,*], Xi Zhou[b]

[a] University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100049, PR China
[b] Intelligent Multimedia Technique Research Center, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, PR China

ARTICLE INFO

Communicated by Dr. Qingshan Liu

Keywords:
Face recognition
Data augmentation
Landmark perturbation
Image synthesis
3D reconstruction

ABSTRACT

Recently, Deep Convolution Neural Networks (DCNNs) have shown outstanding performance in face recognition. However, the supervised training process of DCNN requires a large number of labeled samples which are expensive and time consuming to collect. In this paper, we propose five data augmentation methods dedicated to face images, including landmark perturbation and four synthesis methods (hairstyles, glasses, poses, illuminations). The proposed methods effectively enlarge the training dataset, which alleviates the impacts of misalignment, pose variance, illumination changes and partial occlusions, as well as the overfitting during training. The performance of each data augmentation method is tested on the Multi-PIE database. Furthermore, comparison of these methods are conducted on LFW, YTF and IJB-A databases. Experimental results show that our proposed methods can greatly improve the face recognition performance.

1. Introduction

Face recognition in unconstrained environment has become increasingly prevalent in many applications, such as identity verification, intelligent visual surveillance and immigration automated clearance system. The classical pipeline of a modern face recognition system typically consists of face detection, face alignment, feature representation, and classification. Among them, feature representation is the most fundamental step. An excellent feature can improve the performance to some degree. Up to now, many approaches of face representation have been proposed. Hand crafted features, such as LBP [1], SIFT [2], were early used to extract image's appearance feature. Later, encoding-based features were developed to learn discriminative feature from data. For example, Fisher vector [3] use unsupervised learning techniques to learn the encoding dictionary from training data. Recently, convolutional neural networks (CNNs) provides a supervised or unsupervised learning framework for robust feature learning, and has demonstrated state-of-the-art performances [4,5].

Since LeNet-5 [6] was firstly proposed by LeCun et al., variant CNNs have been designed and are prevalent in image classification [7,8] and object detection [9]. They also have brought a revolution in face recognition, and even outperform human recognition performance [10,11,5]. For example, DeepID3 [10], FaceNet [11], BAIDU [5], have reached over 99% face verification accuracy on the widely used Labeled Faces in the Wild (LFW) database [12].

In order to achieve better performance, the networks become much deeper and wider [13]. Therefore, directly training a deep network from scratch requires a large amount of labeled face images, because there are many parameters in a deep network. Sometimes, training with limited data will easily leads to overfitting. With large network and limited training data the test error keeps increasing after several epochs even though the training error is still decreasing as the training epoch increased [14]. In order to address this problem, a large number of strategies have been proposed: fine-tuning models trained from other large public databases (e.g., ImageNet [15]), adopting various regularization methods(e.g., Dropout [14], Maxout [16], and DropConnect [17]), collecting more training data [18,4,11]. At present, collecting more training data is directly way to improve the performance. With more training data, the trained model has stronger generalization ability. Many state-of-the-art methods are based on large scale training datasets. For instance, DeepFace [4] trained on 4 Million photos of 4 k people; FaceNet [11] trained on 200 Million photos of 8 Million people.

By taking great advantage of social networks on Internet, a large number of images, including faces, objects, scenes, can be easily crawled by search engines. Being able to access large amount of data meets the needs of deep learning training, but annotating data is a tedious, laborious, and time-consuming work, which even requires volunteers with specific expert knowledge. As size of dataset increasing, mistakes, such as wrong labeling, redundancy and duplication are

* Corresponding author.
E-mail addresses: lvjiangjing@cigit.ac.cn (J.-J. Lv), shaoxiaohu@cigit.ac.cn (X.-H. Shao), huangjiashui@cigit.ac.cn (J.-S. Huang), zhouxiangdong@cigit.ac.cn (X.-D. Zhou), zhouxi@cigit.ac.cn (X. Zhou).

http://dx.doi.org/10.1016/j.neucom.2016.12.025
Received 7 March 2016; Received in revised form 25 July 2016; Accepted 7 December 2016
Available online 08 December 2016
0925-2312/

inevitable. Needless to say, getting a large scale database with correctly labeled is too difficult and expensive for research groups, particularly in academia. Therefore, data augmentation methods have been emerged to generate large number of training data using label-preserving transformations, such as flipping and cropping [7,19], color casting [20], blur [21], etc. Experiments in [19] have shown that flipping and cropping reduced the top-1 error rate by over 2% in the ILSVRC-2013. Color casting, blur and contrast transformations, help the trained model equipped with a strong generalization ability to unseen but similar noise patterns in the training data [7,20,21].

However, the above mentioned methods, which can be efficient to improve neural network based image classification systems for different circumstances, are still not enough for face images. Face image has its own particularity and the main challenges for face recognition including poses, illumination, occlusion, etc. The previous common used data augmentation methods, which just make some simple transformations, cannot handle these problems. Hence, face specified data augmentation methods have been proposed. Jiang et al. [22] proposed an efficient 3D reconstruction method to generate face images with different poses, illuminations and expressions. Mohammadzade and Hatzinakos [23] proposed an expression subspace projection method to synthesize new expression images for each person. Seyyedsalehi et al. [24] tried to generate visual face images with different expressions by using nonlinear manifold separator neural network (NMSNN). Most of previous methods are suitable to constrained environment and only generate fixed types visual face images.

As various poses, illumination and occlusion are common problems in face recognition, these factors not only influence face image preprocessing such as face alignment but also affect face image feature extraction. Meanwhile, the training dataset of face recognition is limited and each person only has a few types of images. Even though DCNNs have a powerful representation ability, they still need different kinds of face images in each subject to learn face variations. At present, the limited training dataset is far from enough for robust feature representation model training and seriously decrease the recognition accuracy in these situations. In this paper, we propose five special data augmentation methods dedicated to these factors: (LP), hairstyles synthesis (HS), glasses synthesis (GS), poses synthesis (PS) and illuminations synthesis (IS). These methods aim to alleviate the impacts of misalignment, pose variance, illumination changes and partial occlusions. Moreover, they can be widely used to unconstrained environment. LP method which randomly perturbs the locations of landmark position before face normalization makes feature extraction model robust to misalignment (e.g., translation, rotation, scaling and shear). HS and GS can generate different hairstyles and glasses giving a face image, which enlarge the training set and make the model robust to similar occlusion. 3D face reconstruction, in contrast to [22], is able to reconstruct 3D face model from image with large pose. When the 3D face model reconstructed, we can use it to imitate different poses and illumination, which make the DCNN model robust to different poses and illuminations. Each data augmentation method is verified on Multi-PIE database. The comparison of different data augmentation methods are conducted on Labeled Faces in the Wild database (LFW) [12], YouTube Faces database (YTF) [25] and IARPA Janus Benchmark A database (IJB-A) [26]. Experimental results show that the proposed data augmentation methods can greatly improve the performance of face recognition.

The rest of this paper is organized as follows. Section 2 reviews related previous works. Our approaches of data augmentation are introduced in Section 3. The experimental results are presented in Section 4 and conclusions are drawn in Section 5.

2. Related work

At present, only a few datasets are publicly available, e.g. CASIA-WebFace dataset [27] including 10,575 subjects and 494,414 images, CACD dataset [28] including 2000 subjects and 163,446 images. Compared to the dataset used by the Internet giants like Google [11], which contains 200 million images and 8 million unique identities, the existing publicly accessible face datasets are relatively small and not enough for large DCNN model training.

Thus, a number of data augmentation methods have been proposed to artificially extend the database. Vincent et al. [29] introduced Gaussian noise, Masking noise and Salt-and-pepper noise to generate more corrupted images for training Stacked Denoising Autoencoders. Howard [19] adopted flipping and cropping to enlarge the training dataset, which is widely used in the following studies [27,30], and Xu et al. [31] even integrated the original face image and its mirror for improving representation-based face recognition performance. Xie et al. [32] added images with Gaussian noise to generate large number of noisy images. A number of methods are introduced by Wu et al. [20], such as color casting which alters the intensities of the RGB channels, vignetting which makes the periphery of an image darker than that of image center, and lens distortion which is a deviation from rectilinear projection caused by the lens of camera. In addition to these common methods, which are suitable for all kinds of images, data augmentation methods specific to face images were also proposed. For example, Jiang et al. [22] proposed an efficient 3D reconstruction method to generate face images with different poses, illuminations and expressions. Mohammadzade and Hatzinakos [23] proposed an expression subspace projection method to synthesize new expression images for each person, through which more accurate estimation of the within-subject variability was obtained. Seyyedsalehi et al. [24] proposed nonlinear manifold separator neural network (NMSNN) to extract expression and identity manifolds for face images. But most of them are complex and addicted to constrained environment.

As the saying goes: “the more you see, the more you know”, it is also true for deep neural networks. As revealed in previous works [7,20,21], data augmentation methods help the trained DCNN model equipped with a strong generalization ability to unseen but similar noise patterns in the training data. Our goal in this paper is to develop several simple and efficient data augmentation methods specific to face images.

Landmark perturbation. Shan et al. [33] first proposed landmark perturbation method to enlarge the training dataset to deal with misalignment, but they only perturbed each face image's eye coordinates with eight-neighbor. As shown in Fig. 1, some misalignmented landmarks are far from the ground truth, eight-neighbor is not enough to model the misalignment situation in practice. According to alignment error satisfies a Gaussian distribution, we use a Gaussian distribution to model the perturbation range. In addition, we adopt other transformations to enrich visual images of each person.

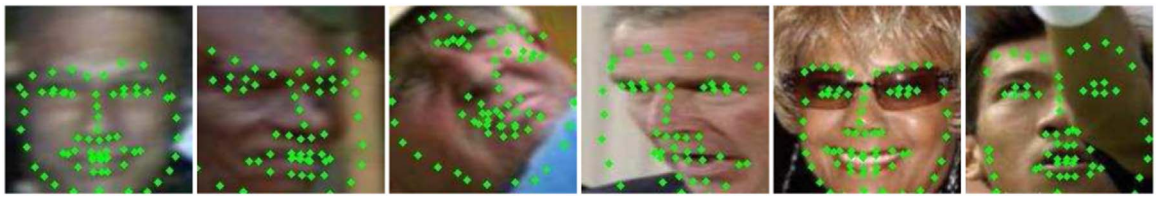


Fig. 1. Examples of landmark misalignment.

Hairstyles synthesis. O'Donnell and Bruce [34] have shown that hairstyle is a particularly important characteristic for recognizing faces. However, in practice, criminals often use different hairstyle masks to cover their hairs or other disguises when committing crimes. Many people especially female change their hairstyles periodically. In addition, due to different hairstyles with different bangs occlude the forehead or even part of eyes, which would affect face recognition performance. In contrast to the previous method [34] which use hairstyles as a positive effect for face recognition, we aim to alleviate the influence of different hairstyles. We collected 87 hairstyle templates with various bangs and adopted a hairstyle synthesis method by referencing to [35]. Contrast to [35] which synthesize hairstyles for makeup applications, we synthesized different images to enrich our training dataset.

Glasses synthesis. Partial face occlusion is one of the most challenging problem in real word face recognition. While eyeglasses and sunglasses which placed in front of the face are the most common occlusion in face. As eyes regions are discriminative for face recognition, most feature extraction methods are densely sampling or setting large weight coefficient for these regions. So occlusion of glasses which affect feature seriously is a challenging task for face recognition. Traditionally, the similarities of occluded faces are conducted based on non-occluded face regions [36,37]. However, these methods only focus to non-occluded regions and sometimes these regions might be essential for recognition [38]. Wright et al. [39] proposed sparse representation based classification (SRC) method which try to recover the occluded parts of face image through the training sample dictionary. Wen et al. [40] proposed structured occlusion coding (SOC) to learn occlusion dictionary which can deal with large occlusion and make greatly improvement in face recognition accuracy. While most previous works operated on the small face databases in the constrained environment and the occlusion regions sometimes needed to be precisely calculated. It is infeasible in practice situation and large scale face recognition. Contrast to previous method, we collected almost 100 different glasses templates and synthesized glasses for each person in the training dataset. DCNN is adopted to learn the within-class scatter and extract glasses-insensitive feature.

Poses synthesis. Pose variance is one of main difficulties for face recognition. In unconstrained environment, frontal images are seldom captured and each person has limited number of images, which makes face recognition systems perform poorly on faces with various pose. Prince et al. [41] proposed a generative model which creates one-to-many mapping from an idealized identity space to the observed data space for large pose face recognition. However, pose types of each image need to be known in advance for correctly selecting mapping matrices. Face frontalization by 3D face reconstruction models [22,42,43] have been proposed for alleviating the effect of pose variations. While the occlusion regions of faces with large poses cannot be precisely reconstructed, which is hardly to improve the face recognition performance. Jiang et al. [22] proposed an efficient 3D reconstruction method to generate face images with different poses, illuminations and expressions. However frontal face images with neutral expression and normal illumination to reconstruction 3D face model is hard to collect for each person. In order to reconstruct a 3D face model from any poses, we make some improvements based on their method, and build a pose-invariant system. Benefitting from the 3D reconstructed model, we can generate variant poses according to non-frontal face image.

Illuminations synthesis. Illumination is one of important issues for face recognition. Due to the changes in lighting conditions, the same face appears differently which decrease the face recognition performance. Even though numbers of illumination normalization methods [44,45] have been proposed, most of them tend to fail under severe illumination variations (e.g., heavy shadows and overexposure) and achieve good performance under constrained conditions. In contrast to these methods, we tend to extract illumination-insensitive feature through DCNN. Hence, each person with different illumination face images are necessary for robust model training. In order to generate variant illumination images of 3D face model, we use the proposed method in [22] to simulate common lighting conditions, while [22] only generated fixed types of illumination and imitated images in CMU-PIE [46]. In our condition, we traverse different types of illumination and generate face images with different unilateral illuminations.

3. Data augmentation

Due to limited training dataset and each person only has a few types of images, it is not sufficient to train a deep and robust DCNN. A reasonable way to enlarge the training dataset size is to use data augmentation. As revealed in previous works [7,20,21], data augmentation methods help the trained DCNN model equipped with a strong generalization ability to unseen but similar noise patterns in the training data. In this Section, we will introduce five data augmentation methods specific to face images.

3.1. Landmark perturbation

Affine transformation is widely used for face normalization, which warps the face image to predefined canonical template using the detected landmarks. With positions of landmarks accurately located, faces can be well warped and each part of face between images have a good correspondence among each other, which is beneficial for feature extraction. But in practical situations, the positions of landmarks cannot always be precisely located because of occlusions, blurs and large poses, as shown in Fig. 1.

This phenomenon which calls curse of misalignment in face recognition was first proposed by Shan et al. [33]. They revealed that imprecise localization of the facial landmarks would abruptly degenerate the face recognition performance. In order to address this problem, they perturbed each face image's eye coordinates from their ground-truth positions in a mode of eight-neighbors deviation and generated totally 81 virtual images. Then the face recognizer was introduced to model the misalignment variations. In contrast to [33], we aim to learn an invariant feature which is robust to misalignment. As it is observed in [33], the alignment error satisfies a Gaussian distribution. Therefore, we use a Gaussian distribution to model the perturbation range instead of eight-neighbors deviation.

$$P_i^* = P_i + r \tag{1}$$

$$r \sim N(\mu, \Sigma) \tag{2}$$

where P_i is the ground-truth position, P_i^* is position after perturb, the perturbation range r is Gaussian distributed and for simplicity the mean is set as 0 and the covariance matrix is a diagonal matrix sharing the same variance σ^2. We generate visual images which are much closer to the practical misalignment situations, and the trained DCNN model has much better generalization ability.

It is not difficult to understand that we can decompose affine transformation matrix into rotation, shearing, scaling and translation transformations and adjust each component's parameters to get different distorted images. But decomposition affine transformation matrix is complex and time-consuming, so we introduce an equivalent way by perturbing the locations of landmarks and use the perturbed landmarks for the following normalization, which is efficient to generate a large number of misalignment images. The process of LP method is as follows.

Firstly, face regions are detected by a Viola-Jones [47] based face detector and resized to a fixed size of 160×160. Secondly, the Supervised Descent Method (SDM) algorithm [48] is adopted to detect the three facial landmarks: the centers of eyes and mouth. Thirdly, we randomly perturb the locations of landmarks by Gaussian distribution

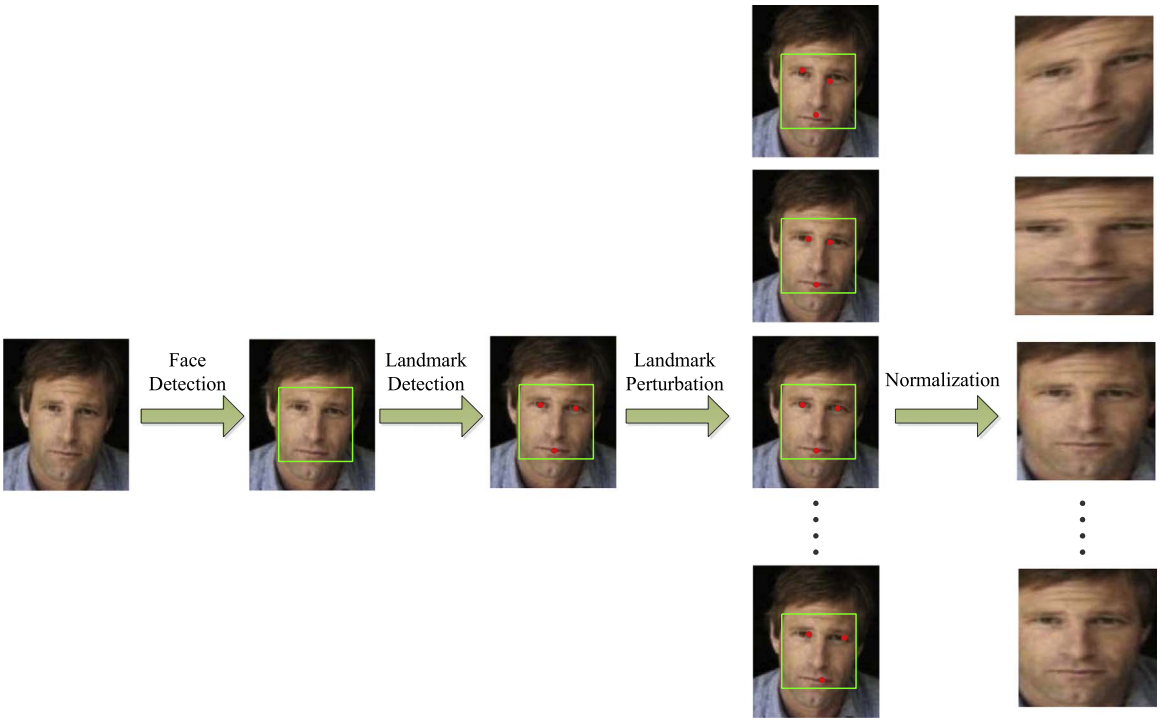


Fig. 2. The process of landmark perturbation.

to generate a series of sets of landmarks. Lastly, face images are normalized by these series of sets of landmarks respectively according to a predefined canonical template.

The pipeline of LP method is shown in Fig. 2. A large number of misaligned face images are generated, including translation, rotation, shearing and scaling. In addition, we also adopt other transformations (e.g., linear conformal transformation, piecewise linear transformation) to enrich our visual images.

3.2. Hairstyles synthesis

In this section, we introduce a hairstyle synthesis method by referencing to [35], which can effectively synthesize different hairstyles for a test face. In face recognition systems, faces are usually cropped from images and most background are removed, excepting bangs which occlude face with different styles. In order to alleviate the effect of different bangs, we synthesize different hairstyles for each training image making model robust to hairstyles. First of all, we collected almost 87 hairstyle templates with various bangs, which are the most common styles in daily life. Then we use the obtained hairstyle templates to synthesize target images, containing two steps: alignment and alpha blending, as shown in Fig. 3.

In the alignment step, a hairstyle template is aligned to the test image according to landmarks. First, the hairstyle template is roughly aligned to the test face by affine transformation. Second, according to the contour points of landmarks as shown in left side of Fig. 5, the middle subset points are obtained by extending on the horizontal direction with half of the interocular distance. Third, combining with 10 points in the edge of the image, whose coordinates are determined by the corner, horizontal lines of eye centers and mouth corners, Delaunay triangulation [49] is used to construct 46 triangles. Lastly, piecewise linear affine transformation [50] is adopted to align the corresponding triangles between the hair template and the test face.

In the alpha blending step, the hair template T and test face I are blended to form a new face image S.

$$S = \alpha I + (1 - \alpha)T \tag{3}$$

where α is a weight parameter, obtained from the hair template. Examples of synthesis results of 87 hairstyles are shown in Fig. 4.

3.3. Glasses synthesis

As wearing glasses is the most common and unavoidable situation in occlusion face recognition, contrast to previous methods [36,39,40], we aim to generate a large number of visual face images with different glasses for each person in training dataset. DCNN is adopted to learn the within-class scatter and extract robust features. In order to make DCNN model robust to different glasses, we collected 100 different glasses templates, which used to synthesize the test image with different glasses. Different from hairs, glasses is a rigid structure, so affine transformation is directly used to align glasses template to the test face image and blend the test image into a new face image. The process is shown in Fig. 5.

First, we use the detected locations of eyes and mouth both in a glasses template and test image. Then according to the locations, we calculate the affine transformation parameters and align glasses template to the test image. Last, by using the mask of the corresponding glasses template, the glasses template is blending to the test image according to Eq. (3). Examples of synthesis results of 100 glasses are shown in Fig. 6.

3.4. 3D face reconstruction

2D-to-3D face reconstruction is to reconstruct a 3D face model from a single face image, and it is used to generate virtual frontal faces recently [22,42,43]. Similar with the work in [22], we first reconstruct a 3D shape model from an input image, and then project the image texture to the 3D geometry to complete the texture. Jiang et al. [22] only reconstructs face model of frontal face image, but fails on non-frontal face. We make some improvements based on their method, and build a pose-invariant system. The framework of 3D face reconstruction is shown in Fig. 7, which mainly consists of three steps: pre-processing, shape reconstruction and texture reconstruction.

(1) Pre-processing

Landmark detection and pose estimation for a 2D face image are very important for the initialization of 3D face reconstruction. If the

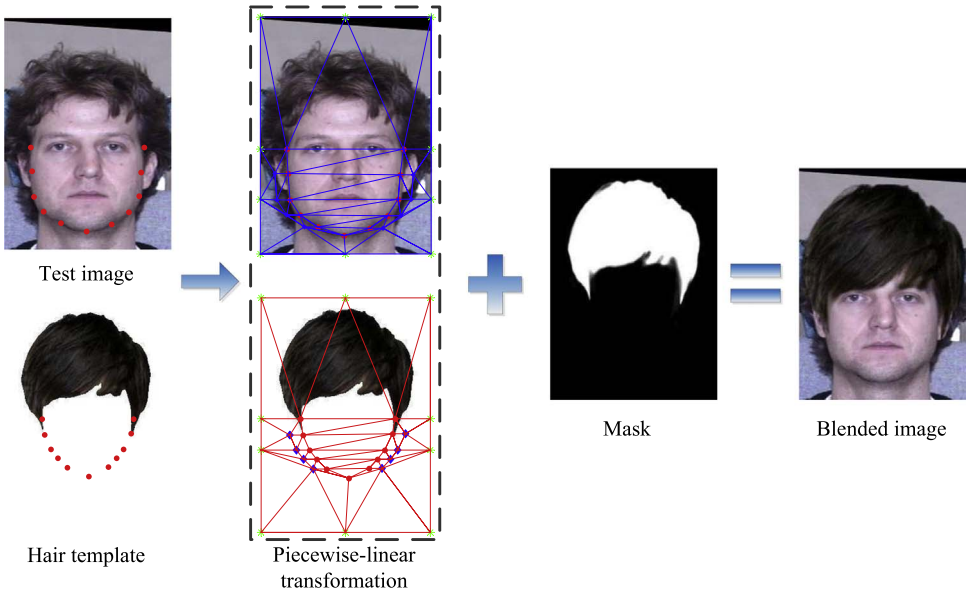


Fig. 3. The process of hairstyle synthesis.

Fig. 4. Examples with different synthesized hairstyles.

location of landmarks and pose of face cannot be correctly provided, the reconstruction inevitably converges to bad local minima [51].

Firstly, faces are detected by a Viola-Jones [47] based face detector and roughly normalized by affine transformation according to the centers of eyes and mouth. Secondly, SDM algorithm [48] is adopted to detect 68 facial landmarks $\{(x_1, y_1), \ldots, (x_{68}, y_{68})\}$. Lastly, according to the 68 landmarks, Delaunay triangulation [49] is used to construct 87 triangles and the pose is estimated by POSIT [52].

(2) Shape reconstruction,

The geometry of a 3D face model can be represented with a shape-vector:

$$S = (x_1, y_1, z_1, \ldots, x_n, y_n, z_n)^T \in R^{3n} \tag{4}$$

where (x_i, y_i, z_i) is the 3D coordinate of the i-th vertice of n vertices. Following the assumption in [22], new shape of 3D face S' can be modeled as a linear combination of the average shape $\overline{S}$ and the principal components $P \in R^{3n\times m}$ with m eigenvectors.

$$S' = \overline{S} + P\boldsymbol{a} \tag{5}$$

where $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)^T \in R^m$ is the coefficient of the shape eigenvectors.

In the landmark detection, t landmarks are selected for 3D reconstruction, where existing correspondences between them and some fixed key vertices of the 3D face model. Let $S_f = (x_1, y_1, x_2, y_2, \ldots, x_t, y_t)^T \in R^{2t}$ be the set of x, y coordinates of those landmarks on the 2D face. It can be assumed that S_f is the sub shape-vector of S. According to Eq. (7), the new 2D face shape S'_f can be expressed as:

$$S_{f'} = \overline{S}_f + P_f\boldsymbol{a} \tag{6}$$

where $\overline{S}_f$ and $P_f \in R^{2t\times m}$ is the x, y coordinates of the feature vertices on $\overline{S}$ and P respectively. The new shape model S' can be reconstructed

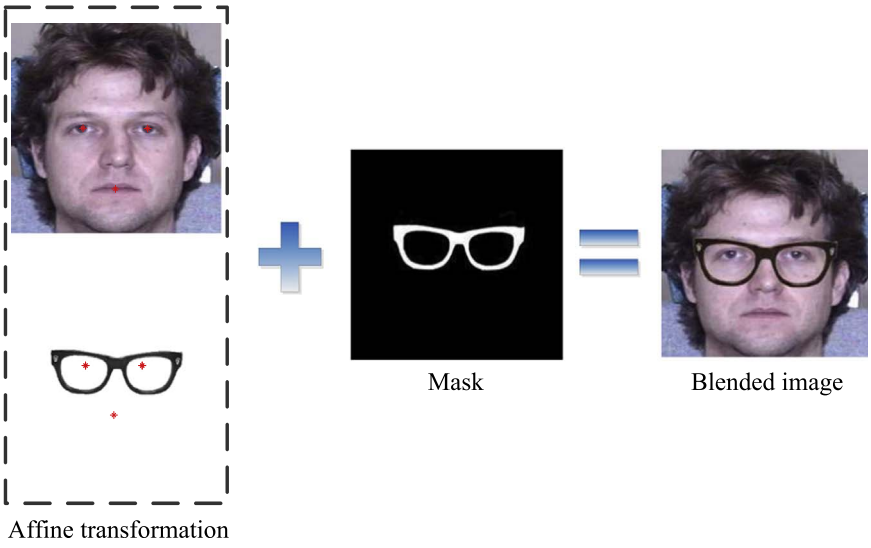


Fig. 5. The process of glasses synthesis.

when geometry coefficient $\boldsymbol{a}$ is obtained. The details of calculating $\boldsymbol{a}$ can be found in [22].

Because the aforementioned shape reconstruction method assumes that input image is a frontal face, there is no pre-processing step before projecting for different face poses. When a non-frontal face appears, S_f should be the face projected from S with the corresponding pose. In order to get a correct S_f, We rotate the 3D model S using pre-estimated pose information.

$$PRS' = PR(\overline{S} + P\boldsymbol{a}) \tag{7}$$

where P is the projection matrix, and R is the rotation matrix calculated form the pose information. Then a new 2D face shape can be expressed as:

$$S'_{f,new} = \overline{S}_f^R + P^R\boldsymbol{a} \tag{8}$$

where $S'_{f,new}$, $\overline{S}_f^R$, and P^R are the rotated vectors. $\boldsymbol{a}$ remains as the same as that in Eq. (6).

(3) Texture reconstruction

After the shape reconstruction, Delaunay triangulation [49] is used to construct 87 triangles in 2D image and we project each triangle to the 3D geometry model. The interpolate method is used to complete the texture reconstruction.

The reconstructed 3D face model is used for the following pose and illumination synthesis.

3.4.1. Poses synthesis

Pose is one of main difficulties for face recognition. In the

Fig. 6. Examples with different synthesized glasses.

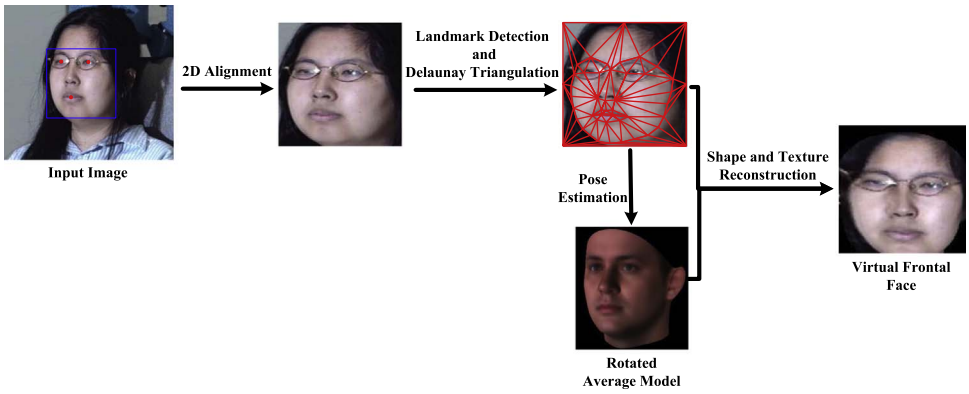


Fig. 7. The process of 3D reconstruction.

Fig. 8. Examples of different poses generate by rotating the reconstructed model.

unconstrained environment, frontal images are seldom captured and each person has limited number of images, which makes face recognition systems perform poorly on faces with various pose. In order to alleviate the influence of poses, we artificially generate more training images with different poses for each image. For each image in the training dataset, we selected the yaw angle between −30° and 30° to 3D face reconstruction. The 3D reconstruction method introduced above is used to get the frontal image and generate multi-view images by rotating the 3D model. The multi-view images generated by 3D model are shown in Fig. 8.

3.4.2. Illuminations synthesis

Illumination is one of important issues for face recognition. Due to the changes in lighting conditions, the same face appears differently which decrease the face recognition performance. Hence, each person face images including different illumination are necessary for robust model training. In our illumination synthesis process, different form [22] we traverse different types of illumination and generate face images different unilateral illuminations.

Two types of lights are applied to the 3D models: environment light and movable spot light. The whole face model illumination is controlled by the environment light. Spot light control the specular areas and shadows, such as ambient, diffuse, which can generate images under different unilateral illuminations. For each input face, spot light moves in both horizontal and vertical directions from left-to-right and top-to-down with fix step. Some visual illumination images are shown in Fig. 9. They are similar to the images generated in real environment.

4. Experiments

4.1. Experiment setup

GoogLeNet [8] is a typical DCNN architecture, whose excellent performance has been shown at ILSVRC-2014 contest. It has 27 layers and introduces inception model to find the optimal local construction. The feature extracted from the trained GoogLeNet model is not only discriminative but also low-dimension and sparsity. Due to these merits, we adopted GoogLeNet as default network to train the models for different data augmentation methods throughout our experiments.

CASIA-WebFace dataset [27], which contains 10,575 subjects of 494,414 images, is used as the basic training set. First of all, all face images in CASIA-WebFace were pre-processed by face detection, landmark detection and face alignment. Then in order to compare different data augmentation methods and show each advantages, each method is used to augment the training set to train the corresponding DCNN model respectively. All images used for DCNN model training are normalized and resized to 128×128.

The performance of each data augmentation method introduced in Section 3 is tested on the CMU Multi-PIE database [46], which is collected in constrained environment and contains different poses and illuminations images for each person. In addition, LFW [12], YTF [25] and IJB-A [26] are used to compare each method in unconstrained environment. For unconstrained face recognition, different evaluation protocols are introduced to fully demonstrate the performance of our methods. Finally, experiments are conducted on data augmentation and manual expanded dataset to verify the efficiency of our methods.

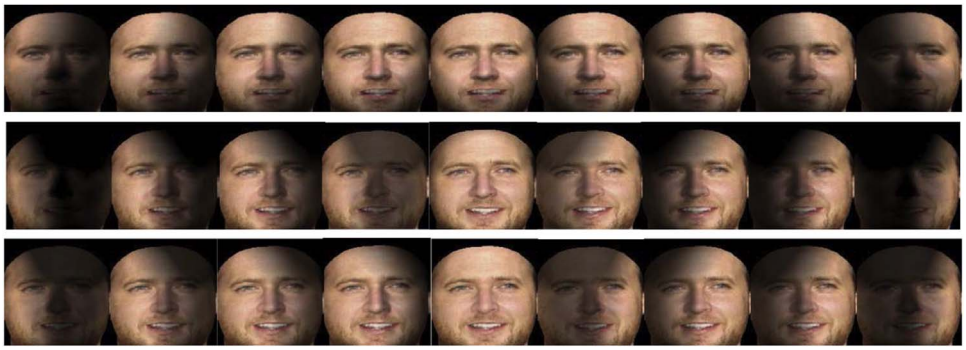


Fig. 9. Examples with different illuminations.

Fig. 10. Examples of artificial misalignment images.

4.2. Experiment in constrained environment

4.2.1. Robustness to misalignment

We evaluate the robustness of LP to misalignments on the CMU Multi-PIE database [46]. The Multi-PIE database contains images of 337 subjects, which are captured in four sessions with simultaneous variations in pose, expression and illumination. All the subjects in Session 1 with 7 frontal images with extreme illumination {0,1,7,13,14,16,18} are used in this experiment.

Images normalized by affine transformation and cropped to 128×128, are denoted as precisely aligned face images. In order to generate misalignment images, an artificial translation of landmarks is introduced to each subject's image. Examples of misalignment images, which contain translation in x direction, y direction, rotation and scale, are shown in Fig. 10. In this experiment, we directly use Euclidean distance for face recognition with neutral expression with illumination {7} as gallery and rest as probe.

The results are shown in Table 1. The **Baseline** is the performance obtained by training the DCNN model with images precisely aligned and testing on images precisely aligned, and **LP** and **No-LP** represent the results obtained by training DCNN with misaligned images generated by landmark perturbation in the DCNN training phrase or not and testing on the artificial misaligned images. We note that images misaligned by the inaccurate locating of landmarks make performance decrease. While using LP method during training, the extracted features are robust to misalignment, which make the performance of **LP** closer to the **Baseline**.

4.2.2. Robustness to different hairstyles

The efficient of hairstyle synthesis data augmentation method is tested on the CMU Multi-PIE database [46]. The DCNN models with adding different synthetic hairstyle images or not are trained respectively. The frontal images in Session 1 with neutral expression and illumination {7} are selected as gallery. Then, the corresponding images in Session 2 with the same scenario are selected as probe, where each image is used to synthesize almost 87 images with different hairstyles.

The results are listed in Table 2, where the **Baseline** shows the results without using data augmentation method and **Hairs Aug** shows the results by adding hairstyle synthesis data augmentation to enrich the training dataset during training. It can be seen that **Baseline** is sensitive to different hairstyles, especially the bangs which occlude forehead and part of eyes. While data augmentation of hairstyle synthesis, the accuracies reach 100%, which shows that the model trained by hairstyle synthesis data augmentation is robust to hairstyle variations.

4.2.3. Robustness to different glasses

In this experiment, we use the same setting as in Section 4.3, except that any subjects who wear glasses are excluded, leaving almost 154 subjects. For each probe image in Session 2 we synthesized 100 images with different glasses.

Table 1
Recognition rates (%) with misalignment on the Multi-PIE database.

Method	Accuracy
Baseline	99.52
No-LP	98.68
LP	99.33

The results are listed in Table 3. From Table 3, we can learn the impact of different glasses on the face recognition and the robustness of the model trained by glasses synthesis data augmentation. The first 78 glasses are normal glasses. The rest 22 glasses are sunglasses which occlude the whole eyes. Among normal glasses, black-rimmed glasses with bold edge are the main factor which decreases the face recognition performance greatly. For sunglasses, with bigger lens, the larger regions of eyes are occluded, the performance decrease dramatically. Comparing **Baseline** with **Glasses Aug** which is the performance of the model trained by adding glasses synthesis data augmentation, we find that the performance of data augmentation method get much better with 100% accuracies in most cases.

4.2.4. Robustness to pose variation

We verify the robustness of our model trained by adding different virtual pose synthesis images to pose variations using the Multi-PIE database [46]. In this experiment, the frontal images with neutral expression and illumination {7} in Session 1 are selected as gallery. The corresponding subjects in Session 2 with illumination {7} and yaw angle {−45°, −30°, −15°, 0°, 15°, 30°, 45°} as probe.

The results on faces with various pose are listed in Table 4, where **Baseline** is the performance of the model trained without any data augmentation and **Pose Aug** is the performance of the model trained with pose synthesis data augmentation. Comparing the performances **Baseline** with **Pose Aug**, we can learn that pose synthesis data augmentation method significantly increase the accuracies.

4.2.5. Robustness to illumination changes

In this section, we evaluate the robustness of our trained model trained by adding different virtual lighting images to illumination variations. In this experiment, we fix pose and expression, only choose images with different illuminations conditions on the CMU Multi-PIE database [46]. The frontal images with illumination {7} in Session 1 are selected as gallery and the corresponding subjects in Session 2 with extreme illumination {0,2,7,13,19} are selected as probe. The DCNN models with using illumination synthesis data augmentation method or not are trained respectively.

The results are listed in Table 5. Comparing the results between **Baseline** and our method, we find that illumination augmentation method can greatly improve the performance especially under extreme illumination.

4.3. Experiment in unconstrained environment

In this section, we test the performance of the models trained with different data augmentation methods in unconstrained environment. The results of different data augmentation methods will be listed in each experiment. The DCNN models of the following data augmentation methods are trained respectively (**A** and **B** are common used data augmentation methods):

- **Baseline**: No data augmentation;
- **A**: Flipping;
- **B**: Contrast+Blur+Noise+Color casting;
- **C**: Landmark perturbation;
- **D**: Synthesis with different hairstyles;
- **E**: Synthesis with different glasses;
- **F**: Synthesis with different poses;
- **G**: Synthesis with different illuminations;

4.3.1. Verification results on LFW

LFW contains 13,233 images with 5749 identities, which is the *de facto* benchmark for automatic face verification. We follow the standard evaluation setting to measure the performance of different data augmentation methods. The standard evaluation protocol provides 10 folds cross validation, each fold contains 300 genuine image

Table 2
Recognition rates (%) of different hairstyles on the Multi-PIE database.

Method															
Baseline	94.74	100	100	100	100	100	100	100	100	96.84	94.74	100	96.84	97.89	100
Hairs Aug	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Method															
Baseline	100	100	98.95	93.68	98.95	97.89	100	98.95	100	88.42	100	86.32	100	98.95	90.53
Hairs Aug	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Method															
Baseline	95.79	100	100	93.68	100	96.84	100	96.84	100	95.79	82.11	98.95	98.95	100	98.95
Hairs Aug	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Method															
Baseline	98.95	98.95	97.89	98.95	98.95	100	100	100	100	100	100	100	100	95.79	88.42
Hairs Aug	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Method															
Baseline	90.53	83.16	100	82.11	90.53	92.63	95.79	97.89	73.68	93.68	98.95	95.79	96.84	92.63	88.42
Hairs Aug	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Method													Acc_mean		
Baseline	93.68	100	88.42	95.79	100	91.58	85.26	91.58	76.84	85.26	72.63	93.68	95.60		
Hairs Aug	100	100	100	100	100	100	100	100	100	100	100	100	100		

Table 3
Recognition rates (%) of different glasses on the Multi-PIE database.

Method															
Baseline	100	100	97.89	100	98.95	98.95	93.68	100	100	100	100	92.63	98.95	95.79	97.89
Glasses Aug	100	100	98.95	100	100	98.95	98.95	100	100	100	98.95	100	98.95	100	100
Method															
Baseline	97.89	100	100	98.95	100	100	95.79	100	94.74	100	100	98.95	98.95	93.68	97.89
Glasses Aug	100	100	100	98.95	100	100	100	100	100	100	100	100	100	100	100
Method															
Baseline	97.89	95.79	93.68	98.95	95.79	76.84	73.68	97.89	96.84	98.95	93.68	100	100	100	100
Glasses Aug	100	98.95	100	100	100	100	98.95	100	98.95	100	100	100	100	100	100
Method															
Baseline	97.89	100	100	100	0.9895	100	100	100	97.89	100	100	98.95	100	98.95	100
Glasses Aug	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Method															
Baseline	98.95	93.68	100	100	100	96.84	100	97.89	96.84	100	100	100	98.95	98.95	96.84
Glasses Aug	100	98.95	100	100	100	100	100	100	100	100	100	100	100	100	100
Method															
Baseline	100	98.95	97.89	85.26	97.89	91.58	83.16	81.05	90.53	92.63	96.84	97.89	91.58	87.37	77.89
Glasses Aug	100	100	100	100	98.95	100	100	97.89	100	100	100	100	100	100	98.95
Method											Acc_mean				
Baseline	86.32	97.89	95.79	95.79	87.37	93.68	92.63	92.63	95.79	90.53	96.42				
Glasses Aug	100	100	100	100	100	100	100	100	100	100	99.85				

Table 4
Recognition rates (%) of different poses on the Multi-PIE database.

Method	−45°	−30°	−15°	0°	15°	30°	45°	Acc_mean
Baseline	92.81	96.38	97.18	97.99	97.38	96.54	92.84	95.88
Pose Aug	93.89	96.95	98.38	98.79	97.99	98.16	93.00	96.50

Table 5
Recognition rates (%) of different illuminations on the Multi-PIE database.

Method	0	2	7	13	19	Acc_mean
Baseline	95.18	95.78	99.40	95.15	91.57	95.42
Illumination Aug	96.99	98.19	99.40	97.58	97.59	97.95

pairs and 300 impostor image pairs. In this experiment, we train the Joint Bayesian model [53] on 9 splits, and test it on the remaining split. The mean accuracy and standard error of 10 folds cross validation are reported.

Table 6 shows the face verification accuracy with different data augmentation methods, as well as the comparison with other state-of-the-art methods on LFW. Comparing all these data augmentation methods with **Baseline**, we find that each data augmentation method improves the face verification performance, especially landmark perturbation reducing the error by almost 50.6%. When concatenating all the features from all the data augmentation models, denoted as

Table 6
Verification rates (%) with various data augmentation methods on LFW database.

Method	Accuracy ± SE
Baseline: No data augmentation	98.18 ± 0.66
A: Flipping	98.43 ± 0.45
B: Contrast+Blur+Noise+Color casting	98.78 ± 0.61
C: Landmark perturbation	99.10 ± 0.37
D: Synthesis with different hairstyles	98.77 ± 0.42
E: Synthesis with different glasses	98.67 ± 0.68
F: Synthesis with different poses	98.30 ± 0.61
G: Synthesis with different illuminations	98.47 ± 0.62
Fusion: Combine all data augmentations	99.33 ± 0.39
DeepFace [4]	97.35 ± 0.25
DeepFace+ [54]	98.37 ± –
DR+Joint Bayse [27]	97.73 ± 0.31
FaceNet [11]	99.63 ± 0.09

Fusion, we achieve 99.33% accuracy, much better than DeepFace [4] and DR+Joint Bayse [27]. Even though FaceNet [11] achieved the best result, it used enormous training dataset which is a thousand time large than ours and adopted triplet loss strategy.

4.3.2. Identification results on LFW

Even though the standard protocol of LFW has been widely used for benchmark evaluation, it only tests the verification rate with 3000 pairs of genuine matches and 3000 pairs of impostor, which does not make full use of all the available data, and a good verification cannot guarantee a good identification. Following the setting in [55], we conduct both closed set and open set identifications on LFW. For convenience, we directly use Euclidean distance to calculate similarities in the following experiments.

(1) Closed set identification

In closed set identification experiments, one frontal face per subject of LFW is selected as gallery and the remaining LFW images as probes. In this experiment, the gallery set consists of 4249 images and the probe set contains 3143 images.

The closed set identification rates of different data augmentation methods are listed in Table 7. We also note that all data augmentation methods increase the accuracy to a degree, such as landmark perturbation achieves a 12% improvement in Rank-1 accuracy over the **Baseline**. Comparing with previous results, our results are much better. When concatenating the features from all the data augmentation models, we achieve more than 90% accuracy.

(2) Open set identification

In the open set identification experiments, the gallery consists of 596 subjects with at least two images in the LFW and we randomly select one image per subject as the genuine gallery set; the probe consists of 10,090 images, including 596 genuine probe images and 9494 impostor probe images. For open-set identification is to determine the identity of the probe or to reject the probe, the performance are reported by taking consideration of the detection and identification rate (DIR), rank of retrieval and the false accept rate (FAR).

Table 8 summarizes the performance of all data augmentation methods, as well as previous results with open-set identification rates at Rank 1 and FAR=1%. These results also show the excellent performance of data augmentation methods. Nevertheless, the best recognition rate is still very low, even though it achieves a good results in verification and closed set identification, indicating that open-set identification is a challenging task.

The ROC curves at rank 1 in Fig. 11(a) show that the open-set identification performance significantly drops with the decreasing value of FAR. What is worse, according to the open-set identification Cumulative Match Characteristic (CMC) curves shown in Fig. 11(b), increasing the number of ranks helps very little in improving the performance. This is because genuine matching scores at lower ranks are hardly be larger than the decision threshold.

Table 7
Closed set identification rates (%) with various data augmentation methods on LFW database. Performance is shown as rank retrieval results at Rank-1, 20, 100, and 200.

Method	R–1	R–20	R–100	R–200
Baseline: No data augmentation	75.66	94.78	98.31	99.24
A: Flipping	81.86	96.53	98.98	99.46
B: Contrast+Blur+Noise+Color casting	82.31	96.66	99.01	99.40
C: Landmark perturbation	87.43	97.77	99.33	99.59
D: Synthesis with different hairstyles	81.58	96.34	98.73	99.20
E: Synthesis with different glasses	86.06	97.65	99.24	99.62
F: Synthesis with different poses	80.78	95.86	98.70	99.11
G: Synthesis with different illuminations	81.10	96.02	98.73	99.30
Fusion: Combine all data augmentations	91.09	98.60	99.55	99.75
COTS-s1 [55]	56.7	78.1	87.1	90.2
COTS-sum [55]	66.5	85.9	92.4	95.1
DeepFace [4]	64.9	–	–	–
DeepFace+ [54]	82.5	–	–	–

Table 8
Open set identification rates (%) with various data augmentation methods on LFW database. Performance is shown with Rank-1 and FAR=1%.

Method	DIR
Baseline: No data augmentation	52.18
A: Flipping	55.20
B: Contrast+Blur+Noise+Color casting	62.08
C: Landmark perturbation	67.28
D: Synthesis with different hairstyles	58.89
E: Synthesis with different glasses	62.92
F: Synthesis with different poses	56.71
G: Synthesis with different illuminations	56.54
Fusion: Combine all data augmentations	71.30
COTS-s1 [55]	25
COTS-sum [55]	35
DeepFace [4]	44.5
DeepFace+ [54]	61.9

4.3.3. Experiment results on YTF

To verify the generalization ability of data augmentation methods, we also test the performance on a video face dataset, YouTube Faces (YTF) [25]. The image quality of YTF is worse than web photos due to motion blur, viewing distance and high compression ratio. YTF consists of 3425 YouTube videos of 1595 subjects which are divided into 5000 video pairs and 10 splits. The video-level face verification performance is measured by mean recognition accuracy. For each pair of training videos, we randomly select one image from each video and label them as same or not-same in accordance with the video pairs. In our experiments, we only use the selected image pairs to train and test.

Table 9 summarizes results of different data augmentation methods. Among all data augmentation methods, the combination of contrast, blur, noise, color casting and landmark perturbation perform best, which achieved over 93% accuracy. By concatenating all data augmentation features, we achieved 94.08%, reducing the error by 26% comparing with **Baseline**. Comparing with previous works, Our results outperform DeepFace [4] and DR+Joint Bayse [27], but slightly lower than FaceNet [11]. FaceNet is benefit for its enormous training dataset, which is a thousand time larger than ours, and triplet loss strategy.

4.3.4. Experiment results on IJB-A

The IJB-A database [26] is a recently released database which aims to push the frontiers of unconstrained face recognition. It contains 500 subjects, with 5712 images and 2085 videos. Contrast to LFW with still images and YTF with video frames, IJB-A includes both images and videos of each person. In addition, it defines both verification and identification (open and closed set) protocols. The evaluation protocol of IJB-A is set-to-set comparisons, rather than using single images comparisons. Thus, a probe may consists of a combination of individual images and videos. In our experiment, we use the average feature in each set to make comparisons.

(1) Face verification

The face verification contains 10-fold cross-validation splits. Each split contains about 11,748 pairs of sets (1756 genuine pairs and 9992 impostor pairs) on average.

Following [26], we summarized our verification performance using True Accept Rates (TAR) at a fixed FAR. The results are shown in Table 10, where TARs at FAR with 0.1, 0.01 and 0.001 are reported. Obviously, our data augmentation methods can significantly improve the face verification performance. When comparing with previous methods, our methods achieve much better performance, especially at lower FAR.

(2) Closed set identification

For closed set face identification, it also contains 10-fold cross-

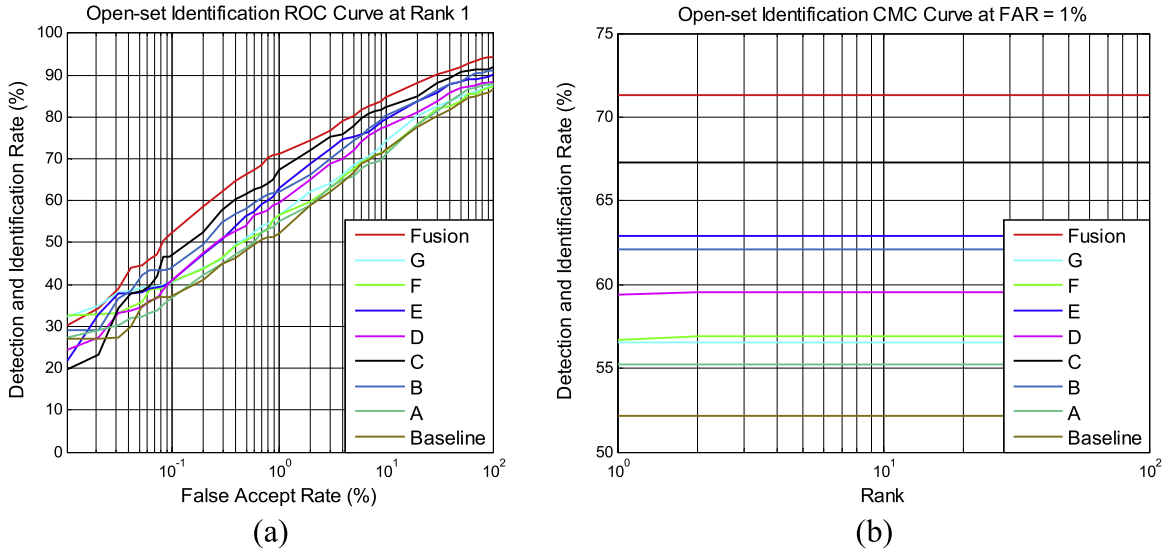


Fig. 11. Open set identification curves. (a) ROC curves at rank 1. (b) CMC curves at FAR=%1.

Table 9
Verification rates (%) with various data augmentation methods on YTF database.

Method	Accuracy ± SE
Baseline: No data augmentation	92.00 ± 1.10
A: Flipping	92.32 ± 1.09
B: Contrast+Blur+Noise+Color casting	93.28 ± 0.98
C: Landmark perturbation	93.22 ± 1.14
D: Synthesis with different hairstyles	92.60 ± 1.19
E: Synthesis with different glasses	92.88 ± 0.73
F: Synthesis with different poses	92.22 ± 1.37
G: Synthesis with different illuminations	92.34 ± 1.73
Fusion: Combine all data augmentations	94.08 ± 1.23
DeepFace [4]	91.4 ± –
DR+Joint Bayse [27]	92.24 ± 1.28
FaceNet [11]	95.12 ± 0.39

Table 10
TARs (%) at 0.1, 0.01 and 0.001 FAR of face verification with various data augmentation methods on IJB-A database.

Method	TAR@FAR		
	0.1	0.01	0.001
Baseline: No data augmentation	89.43	73.80	55.83
A: Flipping	90.72	75.65	57.22
B: Contrast+Blur+Noise+Color casting	91.40	77.32	59.56
C: Landmark perturbation	92.59	77.92	60.69
D: Synthesis with different hairstyles	90.54	74.47	56.10
E: Synthesis with different glasses	90.87	76.52	59.49
F: Synthesis with different poses	89.24	73.92	52.98
G: Synthesis with different illuminations	91.55	75.58	56.18
Fusion: Combine all data augmentations	93.60	78.98	62.89
OpenBR [56]	43.3	23.6	10.4
GOTS [26]	62.7	40.6	19.8
Sparse ConvNet [57]	92.7	72.6	46.0

validation splits. For each split, there are about 1187 genuine probe sets and 112 gallery sets on average.

The performance is evaluated based on genuine samples retrieved at or below a specific rank. Experiment results are listed in Table 11, where Rank 1 and Rank 5 accuracy are reported. Due to face identification is a much difficult task than face verification, the identification rates are relatively lower than verification. While with introducing data augmentation methods, we improve identification rates to a degree. Besides, our methods perform significant better than previous methods.

Table 11
Closed set identification rates (%) with various data augmentation methods on IJB-A database.

Method	Rank−1	Rank−5
Baseline: No data augmentation	79.66	87.90
A: Flipping	80.80	88.86
B: Contrast+Blur+Noise+Color casting	82.17	90.02
C: Landmark perturbation	83.22	90.91
D: Synthesis with different hairstyles	80.97	88.64
E: Synthesis with different glasses	81.80	88.95
F: Synthesis with different poses	81.70	90.12
G: Synthesis with different illuminations	81.76	89.85
Fusion: Combine all data augmentations	83.91	91.01
OpenBR [56]	24.6	37.5
GOTS [26]	44.3	59.5

(3) Open set identification

The protocol of open set identification is similar to closed set face identification except that it added about 576 impostor probe sets in each split.

Following the advice of [26], the performance is evaluated using the false positive identification rate (FPIR) and the false negative identification rate (FNIR). FPIR measures the fraction of impostor probe images accepted at a given threshold, while FNIR measures the fraction of genuine probe images rejected at the same threshold. The results are listed in Table 12, where the FNIRs at FNIR with 0.1 and 0.01 are reported. It can be observed that data augmentation methods help significantly reduce FNIR. When combining all data augmentation methods, FNIRs are reduced to 26.04% at FPIR=0.1 and 41.39% at FPIR=0.01.

4.4. Comparison between data augmentation and manual dataset expansion

In this experiment, we have expanded the CASIA-WebFace dataset [27] to an almost twice larger one which contains 25,580 subjects, 2,545,659 face images. Then we compare the performance between data augmentation based dataset expansion and manual dataset expansion. Due to LP performs the best among all these data augmentation methods, we take it for comparison.

We perform the experiments on LFW and YTF databases. The results are listed in Table 13, where **Baseline** and **Manual ex-**

Table 12
FNIRs (%) at 0.1 and 0.01 FPIR of open set identification with various data augmentation methods on IJB-A database.

Method	FNIR@FPIR	
	0.1	0.01
Baseline: No data augmentation	32.76	50.28
A: Flipping	31.18	47.10
B: Contrast+Blur+Noise+Color casting	28.67	44.29
C: Landmark perturbation	27.96	44.70
D: Synthesis with different hairstyles	30.26	47.93
E: Synthesis with different glasses	29.60	45.04
F: Synthesis with different poses	32.57	50.93
G: Synthesis with different illuminations	30.57	48.19
Fusion: Combine all data augmentations	26.04	41.39
OpenBR [56]	85.1	93.4
GOTS [26]	76.5	95.3

Table 13
The face recognition rates (%) between manual and artificial dataset size expansion.

Method	LFW			YTF
	Verification	Rank=1	DIR@FAR=1%	Verification
Baseline	98.18	75.66	52.18	92.00
Manual expanded	98.90	87.18	68.12	92.92
LP	99.10	87.43	67.28	93.22
Fusion	99.33	91.09	71.30	94.08

panded are the performance of the models trained on the CASIA-WebFace dataset and expanded CASIA-WebFace dataset respectively. We note that **LP** gets the similar result with expanded database. When combining all data augmentation methods (**Fusion**), we achieve a much better performance.

5. Conclusion

This paper presents five data augmentation methods for improving face recognition performance, which aim at increasing the effective size of the training set. Compared with previous data augmentation methods, our methods are dedicated to face images and are more efficient in various situations. Experimental results on Multi-PIE database confirm the effectiveness of our methods. Experimental results on popular LFW, YTF and IJB-A databases show that our methods can significantly improve the accuracy of face recognition rate in unconstrained environment. Besides, we compare data augmentation methods with manual dataset expansion which also verifies the efficiency of our data augmentation methods. In particular, giving limited training data, we get the best performance by fusing features from each data augmentation method. In summary, we demonstrate that our data augmentation methods are more efficient and practical, moreover, they are easy to be implemented and integrated.

Acknowledgments

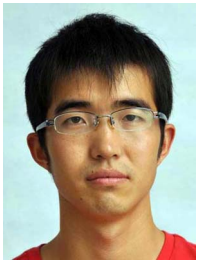
This work was supported by the National Natural Science Foundation of China (NSFC) (Grant no. 61472386), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA 06040103), and Chongqing Research Program of Basic Research and Frontier Technology-No. cstc2016jcyjA0011. The authors would like to thank You-Ji Feng and Cheng Cheng, for valuable discussions.

References

[1] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.
[2] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the IEEE International Conference on Computer Vision, Vol. 2, 1999, pp. 1150–1157.
[3] K. Simonyan, O.M. Parkhi, A. Vedaldi, A. Zisserman, Fisher vector faces in the wild, in: Proceedings of the British Machine Vision Conference, Vol. 5, 2013, p. 11.
[4] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.
[5] J. Liu, Y. Deng, C. Huang, Targeting Ultimate Accuracy: Face Recognition via Deep Embedding, arXiv preprint arxiv:hepth/1506.07310.
[6] 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.
[7] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[9] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[10] Y. Sun, D. Liang, X. Wang, X. Tang, Deepid3: Face Recognition with Very Deep Neural Networks, arXiv preprint arxiv:hepth/1502.00873.
[11] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[12] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.
[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[14] 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:hepth/1207.0580.
[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[16] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, in: Proceedings of the International Conference on Machine Learning, 2013, pp. 1319–1327.
[17] L. Wan, M. Zeiler, S. Zhang, Y.L. Cun, R. Fergus, Regularization of neural networks using dropconnect, in: Proceedings of the International Conference on Machine Learning, 2013, pp. 1058–1066.
[18] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10, 000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891–1898.
[19] A.G. Howard, Some Improvements on Deep Convolutional Neural Network Based Image Classification, arXiv preprint arxiv:hepth/1312.5402.
[20] R. Wu, S. Yan, Y. Shan, Q. Dang, G. Sun, Deep Image: Scaling up Image Recognition, arXiv preprint arxiv:hepth/1501.02876.
[21] T. Ahonen, E. Rahtu, V. Ojansivu, J. Heikkila, Recognition of blurred faces using local phase quantization, in: Proceedings of the International Conference on Pattern Recognition, 2008, pp. 1–4.
[22] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, W. Gao, Efficient 3d reconstruction for face recognition, Pattern Recognit. 38 (6) (2005) 787–798.
[23] H. Mohammadzade, D. Hatzinakos, Projection into expression subspaces for face recognition from single sample per person, IEEE Trans. Affect. Comput. 4 (1) (2013) 69–82.
[24] S.Z. Seyyedsalehi, S.A. Seyyedsalehi, Simultaneous learning of nonlinear manifolds based on the bottleneck neural network, Neural Process. Lett. 40 (2) (2014) 191–209.
[25] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 529–534.
[26] B.F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, A.K. Jain, Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1931–1939.
[27] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning Face Representation from Scratch, arXiv preprint arxiv:hepth/1411.7923.
[28] B.-C. Chen, C.-S. Chen, W.H. Hsu, Cross-age reference coding for age-invariant face recognition and retrieval, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 768–783.
[29] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[30] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[31] Y. Xu, X. Li, J. Yang, D. Zhang, Integrate the original face image and its mirror image for face recognition, Neurocomputing 131 (2014) 191–199.
[32] J. Xie, L. Xu, E. Chen, Image denoising and inpainting with deep neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 341–349.

[33] 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: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 314–320.
[34] C. O'Donnell, V. Bruce, Familiarisation with faces selectively enhances sensitivity to changes made to the eyes, Perception 30 (6) (2001) 755–764.
[35] L. Liu, J. Xing, S. Liu, H. Xu, X. Zhou, S. Yan, Wow! you are so beautiful today!, ACM Trans. Multimed. Comput. Commun. Appl. 11 (1s) (2014) 20.
[36] H. Jia, A.M. Martinez, Support vector machines in face recognition with occlusions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 136–141.
[37] X. Tan, S. Chen, Z.-H. Zhou, J. Liu, Face recognition under occlusions and variant expressions with partial similarity, IEEE Trans. Inf. Forensics Secur. 4 (2) (2009) 217–230.
[38] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, Y. Ma, Face recognition with contiguous occlusion using markov random fields, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1050–1057.
[39] 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.
[40] Y. Wen, W. Liu, M. Yang, Y. Fu, Y. Xiang, R. Hu, Structured occlusion coding for robust face recognition, Neurocomputing 178 (2016) 11–24.
[41] S.J. Prince, J. Warrell, J.H. Elder, F.M. Felisberti, Tied factor analysis for face recognition across large pose differences, IEEE Trans. Pattern Anal. Mach. Intell. 30 (6) (2008) 970–984.
[42] J. Roth, Y. Tong, X. Liu, Unconstrained 3d face reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2606–2615.
[43] T. Hassner, S. Harel, E. Paz, R. Enbar, Effective face frontalization in unconstrained images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4295–4304.
[44] H. Han, S. Shan, X. Chen, W. Gao, A comparative study on illumination preprocessing in face recognition, Pattern Recognit. 46 (6) (2013) 1691–1699.
[45] M.R. Faraji, X. Qi, Face recognition under varying illuminations using logarithmic fractal dimension-based complete eight local directional patterns, Neurocomputing 199 (2016) 16–30.
[46] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis. Comput. 28 (5) (2010) 807–813.
[47] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2001, pp. I–511.
[48] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.
[49] M. De Berg, M. Van Kreveld, M. Overmars, O.C. Schwarzkopf, Computational geometry, in: Proceedings of the Computational Geometry, Springer, 2000, pp. 1–17.
[50] A. Goshtasby, Piecewise linear mapping functions for image registration, Pattern Recognit. 19 (6) (1986) 459–466.
[51] A. Asthana, T.K. Marks, M.J. Jones, K.H. Tieu, M. Rohith, Fully automatic pose-invariant face recognition via 3d pose normalization, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2011, pp. 937–944.
[52] D.F. Dementhon, L.S. Davis, Model-based object pose in 25 lines of code, Int. J. Comput. Vis. 15 (1–2) (1995) 123–141.
[53] D. Chen, X. Cao, F. Wen, J. Sun, Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3025–3032.
[54] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Web-scale training for face identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2746–2754.
[55] L. Best-Rowden, H. Han, C. Otto, B.F. Klare, A.K. Jain, Unconstrained face recognition: identifying a person of interest from a media collection, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2144–2157.
[56] J.C. Klontz, B.F. Klare, S. Klum, A.K. Jain, M.J. Burge, Open source biometric recognition, in: Proceedings of the International Conference on Biometrics: Theory, Applications and Systems, 2013, pp. 1–8.
[57] Y. Sun, X. Wang, X. Tang, Sparsifying neural network connections for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

Jiang-Jing Lv received the B.S. degree in information and computing science from University of Science and Technology of Hunan, Hunan, China, in 2012. He is currently pursuing a Ph.D. degree in pattern recognition at Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China. His research interests include face recognition and deep learning.



Xiao-Hu Shao received the B.E. degree in Telecommunication Engineering from China University of Geosciences in 2009. He received the M.E. degree in Signal and Information Processing from University of Electronic Science and Technology of China in 2012. He joined in Chongqing Institute of Green and Intelligent Technology (CIGIT), Chinese Academy of Sciences as a research trainees from 2012 to 2015. He is pursuing Ph.D. degree in CIGIT and supervised by professor Xi Zhou. His research interests include object detection, 3D face reconstruction and face recognition.

Jia-Shui Huang received the M.S. and Ph.D. degrees in Computer Science from the Zhejiang University, Zhejiang, China, in 2006 and 2010 respectively. He is currently an associate professor at Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences. His research interests include computer vision and machine learning, with focus on face recognition and deep learning.

Xiang-Dong Zhou is an associate professor at the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences. He received the B.S. degree in Applied Mathematics, the M.S. degree in Management Science and Engineering from National University of Defense Technology, Changsha, China, the Ph.D. degree in pattern recognition and artificial intelligence from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1998, 2003 and 2009, respectively. He was a postdoctoral fellow at Tokyo University of Agriculture and Technology from March 2009 to March 2011. From May 2011 to October 2013, he was a research assistant and later an associate professor at the Institute of Software, Chinese Academy of Sciences. His research interests include machine learning and pattern recognition.