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.

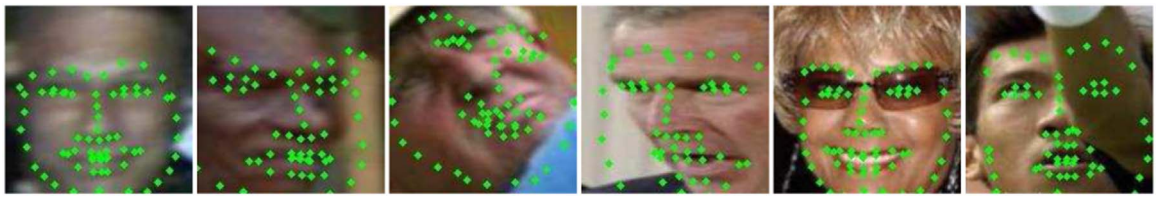

Fig. 1. Examples of landmark misalignment.