Deep Convolution Neural Networks for Automatic Eyeglasses Removal

MAO LIANG, YUEJU XUE, KUNNAN XUE and AQING YANG

ABSTRACT

The facial image under eyeglasses occlusion can degrade face recognition performance. Inspired by the success of deep convolutional neural networks (DCNN) on super resolution, in this paper, a method based on deep convolutional neural network is developed for automatic eyeglasses removal from frontal facial images. To remove eyeglasses on facial images, the proposed approach applied deep convolution neural networks (end-to-end DCNN) to reconstruct the eyeglasses region. We adopt the deep convolutional neural networks (DCNN) approach is designed and trained to learn the mapping between pairs of face images with or without eyeglasses from a large face database in video surveillance. The extensive experiments show that the proposed algorithm can effectively remove eyeglasses, and also can keep the stability of face recognition under eyeglasses on occlusion.

KEYWORDS
Eyeglasses Removal, Deep Convolutional Neural Network, Face Recognition.

INTRODUCTION

Recently, deep learning has shown impressive results on both high-level and low-level vision problems. Face recognition has been one of the most active research areas in pattern recognition and computer vision for its wide potential applications. Many methods have been proposed in the past decades, and some of them have been successfully applied to the practical task of face recognition. In particular, the Super Resolution Convolutional Neural Network (SRCNN) proposed by Dong [6] shows the great potential of an end-to-end DCN in image super-resolution.

Despite a significant level of maturity and several practical successes, face recognition is still a highly challenging task. Several factors affect face recognition performance including pose variations, illumination changes, and most importantly, occlusions [2,4]. Especially, among the last category, eyeglasses are the most common occluding objects, which have a significant effect on the performance of face recognition systems.

Mao Liang, Kunnan Xue, maoliangscau@163.com, College of Electronic Engineering, South China Agricultural University, Guangzhou 510642, China;
Yueju Xue, xueyueju@163.com, College of Electronic Engineering, South China Agricultural University, Guangzhou 510642, China; Guangdong Engineering Research Center for Datamation of Modern Pig Production, Guangzhou 510642, China;
Aqing Yang, Guangdong Engineering Research Center for Monitoring Agricultural Information, Guangzhou 510642, China
Some approaches have been proposed to extract and remove eyeglasses from facial images. In [1], the method studied the statistical mapping between facial images with eyeglasses and their counterparts without eyeglasses. However, for some eyeglasses with no frame, it is difficult to locate the eyeglasses parts accurately. Du [2] used a recursive error compensation approach based on PCA reconstruction by synthesizing eyeglasses and no eyeglasses facial images to remove eyeglasses on facial images. But the method appears too sensitive to changes in lighting conditions. LiYi [3] applied the sparse representation (SR) technique to remove eyeglass on NIR face recognition. The method appears to be very dependent on priori of eyeglasses. Cheng [4] proposed a fatigue detection method with eyeglasses removal which detected object by the OpenCV library and tracked by using a template matching method. However, the method may fail when the rotation of the driver’s head is over 30 degrees. Wong [5] proposed a novel visible information aided eyeglasses removing algorithm for thermal face image reconstruction. The algorithm can significantly elevate the recognition performance. Fernández [6] used Robust Local Binary Pattern and robust alignment to remove glasses on real face images. However, by using these methods above, recognition performance of facial images with different types of glasses (sport glasses, sunglasses, etc) will degrade seriously.

Furthermore, the Super Resolution Convolutional Neural Network (SRCNN) proposed by Dong [6] shows the great potential of an end-to-end DCNN in image super-resolution. The SRCNN has several appealing properties. First, the network directly learns an end-to-end mapping between low- and high-resolution images, with little pre/post-processing beyond the optimization. Second, the SRCNN demonstrate that deep learning is useful in the classical computer vision problem of super-resolution, and can achieve good quality and speed. Inspire by SRCNN, in this paper, we propose an automatic eyeglasses removal method based on deep convolution neural networks (DCNN). First, the DCNN architecture is designed for the eyeglasses region reconstruction, and trained by using selected frontal facial image. The second DCNN architecture is designed for various illumination conditions, facial expressions, and various eyeglasses in face. The proposed architecture learned the mapping between pairs of face images with or without eyeglasses from a large face database in video surveillance containing 19,997 persons.

This paper is organized as follows: Section2 presents the deep convolutional neural network for eyeglasses removal and develops a training algorithm. Section3 provides an implementation procedure for our approach, analyzes the result of the proposed approach. Finally, Section4 concludes the work of this paper.

**PROPOSED APPROACH**

**Deep Convolutional Neural Network for Eyeglasses Removal**

The DCNN has also been successfully applied to solve the high-level and low-level vision problems. In particular, the Super-Resolution Convolutional Neural Network (SRCNN) [7] proposed by shows the great potential of an end-to-end DCNN in image super-resolution. Moreover, the Artifacts Reduction Convolutional Neural Networks (AR-CNN) has been proposed to improve SRCNN by embedding one or more “feature enhancement” layers after the first layer to clean the noisy features [8].
Consequently, by using the current successful SRCNN, we proposed an eyeglasses removal DCNN.

Our eyeglasses removal DCNN is a convolutional neural network that directly learns the mapping between wearing eyeglasses images and no wearing eyeglasses images end-to-end. The proposed DCNN contains three convolutional layers, namely the feature extraction, non-linear mapping and reconstruction layer. The first layer extracts a high-dimensional feature vector from overlapping patches of the wearing eyeglasses image. Then the second layer maps each high-dimensional vector to another high-dimensional vector. At last, the last layer aggregates the patch-wise representations to generate the final no wearing eyeglasses images.

Let us denote the wearing eyeglasses images as $Y$. Our goal is to recover an image $F(Y)$ from $Y$, and we wish that $F(Y)$ is as similar as possible to the ground truth no wearing eyeglasses image $X$. The proposed network can be expressed as:

$$F(Y) = Y$$  \hspace{1cm} (1)

$$F_i(Y) = PReLU(W_i*Y + B_i), i \in \{1, 2\}$$  \hspace{1cm} (2)

$$F(Y) = W_3*Y + B_3$$  \hspace{1cm} (3)

where $W_i$ and $B_i$ represent the filters and biases of the $i$th layer respectively, $F_i$ is the $i$th output feature maps and “*” denotes the convolution operation. The $W_i$ contains $n_i$ filters of support $n_i \times f_i \times f_i$, where $f_i$ is the spatial support of a filter, $n_i$ is the number of filters, and $n_i$ is the number of channels in the input image. Because there is no pooling or full-connected layers in SRCNN[1], the final output $F(Y)$ is of the same size as the input image.

We propose a DCNN architecture that is different from SRCNN[7]. Our DCNN adopts Parametric Rectified Linear Unit(PReLU)[8] as the activation function. The general activation function is presented as:

$$PRelu(x_c) = \max(x_c, 0) + a_c \cdot \min(0, x_c)$$  \hspace{1cm} (4)

where $x_c$ is the input signal of the activation $f$ on the $j$th channel, and $a_c$ is the coefficient of the negative part. We choose PReLU mainly to avoid the “dead features” caused by zero gradients in ReLU[8].

A deeper SRCNN imposes more non-linearity in the mapping stage, which equals to adopting a more robust regress or between the low-level features and the final output[8]. Similarly, our method extracts low-level features by a single layer. This leads to better performance than original SRCNN.

**Training.**

Learning the end-to-end mapping function requires the estimation of network parameters. This is achieved through minimizing the loss between the reconstructed
images and the corresponding ground truth no wearing eyeglasses images. Given a set of no wearing eyeglasses images and their corresponding wearing eyeglasses images, we use Mean Squared Error (MSE) as the loss function:

\[
L(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \| F(Y_i; \Theta) - X_i \|^2
\]  

Where \( n \) is the number of training samples. The loss is minimized using stochastic gradient descent (SGD) with the standard back propagation. We adopt a batch-mode learning method with a batch size of 64.

EXPERIMENTS

In the section, we first collected a mass of wearing eyeglasses and no wearing eyeglasses images to set up the new training and testing dataset. In addition to the new dataset, we then designed different network architectures, and studied the optimal performance factors including the depth and the number of filters. Finally, our approach was applied on the color RGB images and its performance was evaluated on testing dataset.

Training and Testing Datasets

As shown in the literature, deep learning generally benefits from big data training. Consequently, we have collected 39,994 face images covering all possible cases of with or without eyeglasses, pose, gender, age, and lighting condition from web and video surveillance scenes. In order to acquire 19997 pairs of with or without wearing eyeglasses, we artificially synthesized eyeglasses on face images.

The testing datasets includes 300 face images from 150 persons in surveillance camera. For each person, the face images with or without wearing eyeglasses were captured respectively. In this paper, 97 pairs of images were selected from 150 persons randomly to demonstrate the performance of the proposed approach. 11 of these images were evaluated in details to highlight the advantages of the proposed approach. While another 86 pairs of images were presented as supplementary images to further support the findings.

For these face images, we positioned and aligned face images by landmarks [9] depicted in Fig.1(a). Specially, we rotated two eye points horizontally which can overcome the pose variations in roll angle. The distance between the midpoint of eyes and the midpoint of mouth, as well as the y axis of mid-point of eyes, will be used for facial image normalization, because the distance between the midpoint of eyes and the midpoint of mouth is relative invariant to pose variations in yaw angle. And then all the face images were normalized to 128x128 pixels in our experiments as shown in Fig.1(b), and Fig.1(c) shows the aligned face images synthesized wearing eyeglasses.
Implementation details

All experiments were run on a PC with i7-6700 3.4GHz CPU and GTX 1080 GPU and used the open source deep learning framework Caffe[2] for training the model in this paper. In the section, we firstly used the SRCNN baseline network, and set $f_1 = 9$, $f_2 = 1$, $f_3 = 5$, $n_1 = 64$ and $n_2 = 32$, denoted as 9-1-5[7]. The network parameters were randomly initialized from a Gaussian distribution with a standard deviation of 0.001. In the training phase, we randomly selected some face images with or without eyeglasses from each identity as the validation set and the other images as the training set. However, the network cannot convergence.

Then, to ensure the eyeglasses removal DCNN convergence, we used the setting with $f_1 = 9$, $f_2 = 9$, $f_3 = 1$, $n_1 = 64$ and $n_2 = 32$, denoted as 9-9-1. Furthermore, we chose PReLU instead of ReLU as the activation function for the network. In the training phase, we observed the training stops after 100 epochs all the time, until no improvement of the cost function. Initial learning rate was set to 0.01 and final learning rate was set to 0.0001 and updated gradually when the improvement of the
cost function is smaller than a given threshold. Finally, in the condition of comprehensive understanding of the problem and careful design of the network structure, we successfully trained an eyeglasses removal DCNN. The view of eyeglasses removal DCNN is shown in Fig.2.

The results positively indicate that eyeglasses removal DCNN performance may be further boosted using a larger training set, but the effect of big data is not as impressive as that shown in high-level vision problems. Our network contains no pooling layer or full-connected layer, thus it is sensitive to the initialization parameters and learning rate. When we go deeper (e.g., 4 or 5 layers), we find it hard to choose appropriate learning rates that guarantee convergence. Even it converges, the network may fall into a bad local minimum, and the learned filters are of less diversity even given enough training time. Why “deeper is not better” is still an open question, which requires investigations to better understand gradients and training dynamics in deep architectures. Therefore, we still adopt three-layer networks in the following experiments.

Eyeglasses Removal Results

In real world, there are many kinds of eyeglasses, with different materials, different colors and different shapes. In the section, according to the type of glass frame, the wearing eyeglasses face images were divided into three categories, namely wearing semi-rimless frame eyeglasses, wearing full frame eyeglasses, wearing rimless eyeglasses as shown in Fig.3(a)-(b). Note that we utilize these different types of wearing eyeglasses face images to build the training dataset and testing dataset in this paper.

Figure 3. (a) wearing semi-rimless frame eyeglasses face images. (b) wearing full eyeglasses face images. (c) wearing rimless frame eyeglasses face images.
Some results with eyeglasses removal are shown in Fig.4. The training samples of artificially synthesized eyeglasses on face images have removed completely by the proposed approach. Then, our approach was applied to testing dataset with real conditions. As shown in Figs.5, the results clearly prove that 99% eyeglasses on face images have been successfully removed using the proposed approach. The proposed approach successfully avoids these problems caused by the inherent appearance variation on eyeglasses frames, the similarity between face skin and eyeglasses parts, different lighting condition and the similarity between eyebrows and edges of eyeglasses. The inherent appearance variation on eyeglasses frames may degrade the performance of eyeglasses removal algorithms [6], and face skin is very similar to the eyeglasses and eyebrows cross edges of eyeglasses creating substantial ambiguities. Furthermore, the results of eyeglasses removal based on DCNN can effectively removes any kind of eyeglasses no matter the eyeglasses cover eyes or not, which can keep the stability of face recognition under occlusion.
SUMMARY

The paper presents the automatic eyeglasses removal based on deep convolutional neural network (DCNN). Our DCNN removes the eyeglasses region and reconstructs this region by selecting one piece from the facial image. Particularly, we adopt a DCNN to learn the mapping between pairs of face images with or without eyeglasses from a large face database in video surveillance. Our approach integrates eyeglasses detection, eyeglasses localization, and eyeglasses removal in [1] into one end-to-end model. Experimental results demonstrate that our DCNN can provide a good solution to eyeglasses removal. Our approach is robust to lighting variation, the inherent appearance variation on eyeglasses frames and the similarity between face skin and eyeglasses.

ACKNOWLEDGEMENTS

This work was supported by the Science and Technology Planning Project of Guangdong Province, China #2014A020208108, #2015A020209148, and Science and Technology Program of Guangzhou of China #201605030013, #201604016122.

REFERENCES