Shallow convolutional neural network for eyeglasses detection in facial images

Conference Paper · September 2017

DOI: 10.1109/CEEC.2017.8101617

5 authors, including:

Arwa Basbrain
University of Essex
8 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Nassr Azeez
University of Essex
5 PUBLICATIONS **13** CITATIONS
SEE PROFILE

John Q Gan
University of Essex
204 PUBLICATIONS **2,582** CITATIONS
SEE PROFILE

Adrian Clark
University of Essex
147 PUBLICATIONS **968** CITATIONS
SEE PROFILE

Shallow Convolutional Neural Network for Eyeglasses Detection in Facial Images

Arwa M Basbrain[1,2], Inas Al-taie[1,3], Nassr Azeez[1,3]
John Q Gan[1], Adrian Clark[1]
[1] School of Computer Science and Electronic Engineering,
University of Essex, Colchester, UK
{amabas, iyyalt, naazee, jqgan, alien} @essex.ac.uk
[2] Faculty of Computing and Information Technology, King Abdul-Aziz University,
Jeddah, KSA
{abasareen@kau.edu.sa}
[3]Faculty of science, computer department, Baghdad university

Abstract — **Automatic eyeglasses detection plays a major role in many facial analysis systems. To improve the robustness of these systems and cope with real-world applications, a high-speed eyeglasses detector that can achieve high accuracy is needed. Recent studies indicate that the features extracted from convolutional neural networks are compelling. Therefore, this paper presents an effective and efficient method for eyeglasses detection in facial images based on extracting deep features from a well-designed shallow convolutional neural network (CNN). The main contribution of this paper is to address the two essential aspects of CNN: (1) the size of the training dataset required and (2) the depth of the network architecture. To this end, we initialize the learning parameters of the shallow CNN by the parameters of a deep CNN which is fine-tuned on a small dataset. The depth of the neural network is then decreased by removing some convolutional layers after testing its performance on the validation dataset. As a result, a significantly more accurate shallow CNN architecture, Shallow-GlassNet, is obtained, which achieves not only high accuracy but also high speed in eyeglasses detection. Evaluation experiments have been conducted on two large unconstrained facial image databases, LFW and Celeb Faces. The results have demonstrated the superior performance of the proposed framework which achieves a mean accuracy of 99.73%.**

Keywords— Eyeglasses Detection; Convolutional Neural Networks; Shallow CNN.

I. INTRODUCTION

The challenges related to human facial analysis systems can be attributed to many appearance and technical factors. Appearance factors are related to the subject's face, such as pose, facial expression, whereas technical factors are related to the clarity and quality of images, such as illumination variations, shadows, image resolution, and the presence of intervening components such as glasses and hands. Eyeglasses are considered as a confounding factor of human facial analysis systems due to reflections and frame occlusion, which cover arguably the most important part of the face, the ocular region. Moreover, the presence of glasses may cause inaccurate classification, especially when the facial analysis system utilises a convolutional neural network in its models [1-3].

To increase the accuracy, several human facial analysis systems have included a glasses/non-glasses classification phase in their frameworks [4, 5], which leads to increased memory consumption and computation time. Therefore, a high-speed glasses and non-glasses image classifier that can be utilised in human facial analysis systems is needed to improve the robustness of the systems and cope with real-world applications.

Most of the existing approaches for eyeglasses detection utilise handcrafted feature extraction methods. Several pattern recognition projects have demonstrated that deep learning features may provide valuable information about the relationships between raw data and learnt features. The convolutional neural network (CNN) has become the most widely used approach in computer vision in recent years. CNN combines feature extraction and classification together. It is an end-to-end model which receives the raw input data and gives the final classification results without any auxiliary process. Without any handcrafted features, CNNs can handle large training samples, and features are automatically learned by the neural networks. So CNNs can be treated as a powerful automatic feature extractor [6].

Motivated by the success of CNN in various computer vision tasks, this paper proposes a shallow CNN with less complex structures that can be utilised in different facial analysis system frameworks without consuming their resources and, furthermore, achieves high accuracy in realtime. The proposed shallow CNN, called Shallow-GlassesNet, consists of just six layers: three convolution layers, two max-pooling layers, and one fully-connected layer.

The rest of this paper is organised as follows. Some related work is reviewed in Section II. The proposed framework and the structure of the Shallow-GlassesNet are introduced in Section III. Experimental setup and databases used are described in Section IV. Experimental results and discussion are presented in Section V. Section VI draws conclusions.

978-1-5386-3007-5/17/$31.00 ©2017 IEEE

II. Related Work

The existing eyeglasses detection methods can be categorised into two approaches: handcrafted feature approach and deep learning approach. Zhong *et al.* [7] have developed a glasses detection and extraction algorithm, in which detection is realised using edge information within a small area defined between the eyes whilst extraction is achieved with a deformable contour, combining edge features and geometrical features. They obtained two false alarms in their test, by falsely detecting the presence of glasses in facial images. With the facial database used in their experiment, 50% of glasses were accurately extracted, 30% of glasses were extracted with satisfactory results, and the remaining 20% were obtained with fair results. Their experimental investigation on their own images, not a public database, showed an overall detection accuracy of 95.5%.

Wu *et al.* [8] presented a novel glasses detection method in which glasses detectors are learned by boosting simple wavelet feature based weak LUT (look-up-table) classifiers. They investigated the performance of their proposed method using Haar and Gabor features by utilising AdaBoost and SVM. Remarkable performance was obtained by Gabor features and AdaBoost on the public database FERET [9], reporting a glasses detection rate of 98.9%. Experimental results show that the boosting methods have better performance than SVMs.

Fernndez *et al.* [10] used Robust Local Binary Pattern and SVM for eyeglasses detection in their method. The proposed method has been tested over the *Labelled Faces in the Wild* database [11], showing accuracy rate of 98.65%. Du *et al.* [3] proposed a new set of Haar-like features to detect glasses more robustly. Using AdaBoost algorithm, their method achieved a detection rate of 95.11% on the face database CAS-PEAL-R.1[12]. Shao *et al.* [2] proposed a deep convolutional neural network called GlassesNet (GNet). They first pre-trained it for face identification and then fine-tuned it as a glasses detection network. They evaluated their method on different databases, achieving accuracies ranging from 95% to 99.4 %. Their experiments on the Multi-PIE database show that the proposed method is strongly robust to various challenging conditions.

Alberto *et al.* [4] proposed a real-time big data architecture in order to collect, maintain and analyse massive volumes of images related to the problem of automatic glasses detection. This architecture can be used for automatic image tagging, related with glasses detection in facial images. Their innovative algorithm is based on Robust Local Binary Pattern and robust alignment. Experimental results demonstrate that a simple yet efficient algorithm can obtain impressive classification accuracy, achieving 98.65% recognition rate on the LFW database. This algorithm has been tested on the FERET database too, achieving 99.89% recognition rate. Experimental results also show that the proposed algorithm is robust under a wide range of lighting conditions and different poses, and can deal with occlusion, which is very common with sunglasses.

Mohammad *et al.* [1] proposed two schemes for prescription eyeglasses detection. The first proposed scheme is not learning based, which uses Viola-Jones to detect Region of Interest (ROI) followed by glass detection, yielding an overall accuracy of 99.0% on the FERET database and 97.9% on the VISOB database. The second scheme is learning-based, which obtained a best overall accuracy of 99.3% on the FERET database and 100% on the VISOB database. Shaoyi *et al.* [3] proposed an accurate glasses detection algorithm for in-plane rotated faces by using a new set of Haar-like features which represent the features of rotated faces.

Zheng [13] proposed a face detection algorithm using projection profile analysis and an eyeglasses detection algorithm using block/region processing plus prior knowledge. Both algorithms were tested with the ASUIR database (142 thermal face images from 71 subjects) in which the ground truths for both face region and eyeglasses region were established manually. All faces were successfully detected by their algorithms and the averaged overlapping ratio with the ground truths is 0.8998. The other algorithm detected 22 of 23 eyeglasses, and the averaged overlapping ratio with the ground truths is 0.7986.

III. Methods

In this section, we describe the architecture of the Shallow-GlassesNet and the proposed pipeline for glasses and non-glasses image classification in detail. The pre-training and fine-tuning processes for the Shallow-GlassesNet are also introduced.

A. Shallow-GlassesNet Architecture

Inspired by GoogleNet [14] architectures, the proposed Shallow-GlassesNet, as shown in Fig. 1, contains six layers: three convolutional and two max-pooling layers followed by a fully-connected layer. The kernel size and stride of Conv1,

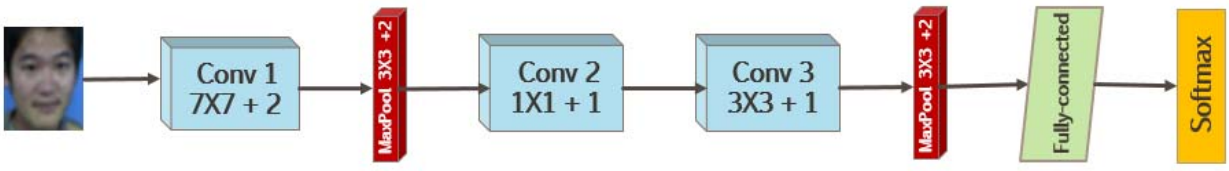

Fig. 1. The Shallow-GlassesNet Structure

Fig. 2. The Eyeglasses Detection Pipeline

Conv2 and Conv3 layers are set as 7×7 (2), 1×1 (1), and 3×3 (1), and their outputs are feature maps of sizes 64, 64 and 192 respectively. Each of these layers has similar corresponding layers in GoogleNet. The three convolutional layers are followed by Rectified Linear Unit (ReLU), which is several times faster than other equivalents with tanh units [14]. Max-pooling is performed over a 3×3 pixel window, with stride 1.

B. *Eyeglasses Detection Pipeline*

As shown in Fig. 2, the proposed pipeline consists of three parts: face detection, features extraction by Shallow-GlassesNet, and classification/detection by SVM. To prepare training data, we used the face detection approach of Viola and Jones [14] and Joint-Face-Detection [15] face detectors. After obtaining and cropping the frontal face region from the visible image, the cropped face image is re-sized to match the input size 224×224 of the Shallow-GlassesNet, and the mean RGB value, computed on the training set, is subtracted from each pixel.

Instead of using our Shallow-GlassesNet to classify the images directly, we use it as a feature extractor. In the second part, the features of the pre-processed image are extracted from the last max-pooling layer of the Shallow-GlassesNet. To increase the accuracy, a Linear Support Vector Machine (SVM) classifier is trained on the extracted features to classify face images with or without egeglasses.

C. *Fine-Tuning and Training Process*

When designing a CNN, the initialization of the network weights is critical since bad initialization can cause gradient instability which could stop learning [16]. To avoid this problem, a GoogleNet was fine-tuned with a small facial database in the first place. The fine-tuned GoogleNet is used as feature extractors. To select the best performance layer according to its accuracy and the time consuming, each layer of the fine-tuned GoogleNet is used to extract features of the validation set images to test its performance. Then, when utilising the Shallow-GlassesNet, we initialised its three convolutional layers with the weights of the corresponding layers of the fine-tuned GoogleNet. As the fully-connected layer is not used for feature extraction, it is initialized randomly.

The initialised weights of Shallow-GlassesNet is kept fixed, which means no fine-tuning is performed. By utilising the Shallow-GlassesNet as feature extractor, the resulting image features are normalised and fed into a linear SVM classifier which is trained on another visible image database of face images with or without eyeglasses.

IV. EXPERIMENTS

A. *Experiment Design*

The USTC-NVIE [17] (NVIE) database is adopted to fine-tune the GoogleNet, train the SVM, and test the proposed pipeline. It contains both posed and spontaneous facial expressions of more than 100 subjects, with illumination for three different directions. The posed database contains the apex expressional images with and without glasses.

The database is divided into two parts: With-Glasses-Dataset and Without-Glasses-Dataset. Fig. 3 shows some sample images from the posed database of seven different facial expressions. The database is small for training the SVM. To deal with the over-fitting problem, we artificially enlarge data set by using different techniques such as horizontal rotation. The original dataset is randomly partitioned into five almost equal sized subsets, without overlapped subject images among them. Of the five subsets, one is retained for testing,

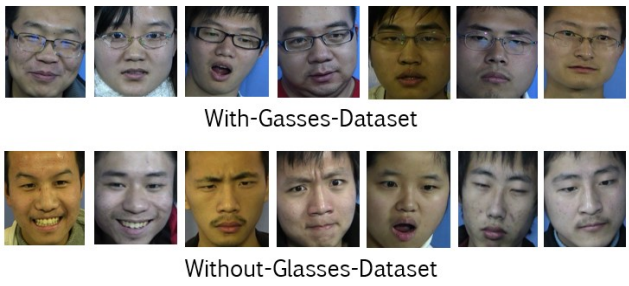

Fig. 3. Samples from With-Glasses-Dataset and Without-Glasses-Dataset from the NVIE posed database for the seven different facial expressions.

and the remaining four are used for fine-tuning, training, and validation. Table 1 illustrates the number of subjects and images in the training, validation, and testing datasets respectively.

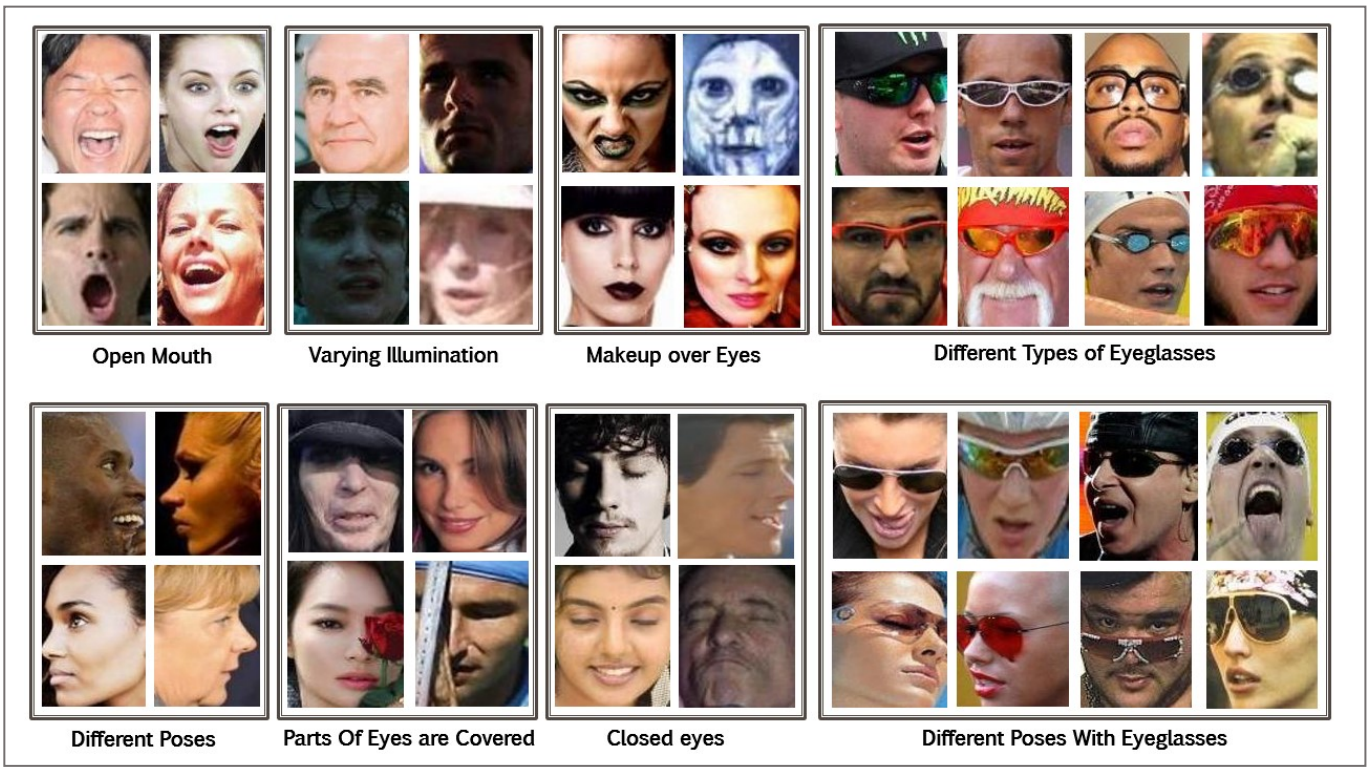

Fig. 4. Samples from the Celebrity Database (Celeba)

TABLE 1. THE NUMBER OF SUBJECTS AND IMAGES IN TRAINING, VALIDATION, AND TESTING DATASETS

NVIE Datasets	Without-Eyeglass		With-Eyeglass	
	Subject	*Image*	*Subject*	*Image*
Train	61	8946	61	8959
Val	20	2813	20	3062
Test	20	3043	20	3002

To evaluate the performance of the proposed methods for eyeglasses detection, we adopt two large facial databases which were created for studying the unconstrained face recognition problem: (1) Labelled Faces in-the-Wild (LFW) [11] and (2) Celeb Faces (Celeba) [5]. The LFW database contains 13,233 face images of 5,749 different people collected from the web, with 1,244 images with eyeglasses. The Celeba database contains 202,599 face images of 10177 different celebrities. Some samples from the Celeba database are shown in Fig. 4 from which it can be seen that eyeglasses detection is a challenging problem.

B. Neural Network Setup

We applied the Caffe toolkit [18] using an NVIDIA GeForce GTX 980 GPU to fine-tune the pre-trained GoogleNet Deep CNN model [14] using the NVIE dataset. GoogleNet was fine-tuned using the stochastic gradient descent with a batch size of 50. The hyper-parameters of the applied training algorithm are as follows: momentum=0.9, weight decay=0.0002, initial learning rate=0.001.

V. RESULTS AND DISCUSSION

Before starting the testing, four-fold cross-validation was adopted to find the extracted features which have the best mean accuracy on the validation datasets. To analyse the efficiency of the Shallow-GlassesNet, we conducted two comparative evaluations to compare the accuracy and speed of the Shallow-GlassesNet and GoogleNet.

Table 2 shows the eyeglasses detection accuracy on validation and testing datasets from the NVIE database by using Shallow-GlassesNet and GoogleNet respectively. To calculate the accuracy of the proposed pipeline, we used the quantity [19] :

$$(TP + TN)/N \quad (1)$$

where TP is the number of true positive detections, TN is the number true negative detections, and N represents the number of face images tested. Table 2 also shows the speed of the Shallow-GlassesNet in comparison with GoogleNet, which indicates that Shallow-GlassesNet is much faster than GoogleNet.

Table 3 shows the confusion matrix and average accuracy of the Shallow-GlassesNet for eyeglasses detection on the LFW database and Celeba database. The results in Table 2 and Table 3 show that the proposed shallow CNN has achieved very high

accuracy. It can be clearly seen from the results that the proposed method is strongly robust to various challenging conditions.

TABLE 2. COMPARISON BETWEEN GOOGLENET AND SHALLOW-GLASSESNET IN TERMS OF ACCURACY AND SPEED

	Accuracy		Speed S/ Image	
	Val	***Test***	***CNN***	***Pipeline***
GoogleNet	90.13	90.92	0.1110	0.1560
Shallow-GlassesNet	99.31	99.78	0.0297	0.0782

TABLE 3. CONFUSION MATRIX AND AVERAGE ACCURACY OF THE SHALLOW-GLASSESNET FOR EYEGLASSES DETECTION

	LFW		Celeba	
	Without	***With***	***Without***	***With***
Without	0.9869	0.0131	0.9604	0.0396
With	0.0273	0.9727	0.0235	0.9765
Accuracy	98%		97%	

VI. CONCLUSIONS

This paper presents an efficient and effective eyeglasses detection framework based on well-designed shallow CNN, called Shallow-GlassesNet. First, the pre-trained GoogleNet was fine-tuned with both eyeglasses and non-eyeglasses images. Then the learned weights of the GoogleNet were copied to the corresponding layers in the Shallow-GlassesNet which was used as feature extractor. A linear SVM was trained on the extracted features to detect eyeglasses. The proposed Shallow-GlassesNet architecture reduced the detection time of eyeglasses detection by roughly a factor of two while retaining high detection accuracy. The main contribution of this paper is designing shallow CNN for eyeglasses detection. Unlike most CNN designs, this shallow design is characterised by high precision and high speed at the same time, making it ideal for use in real-time applications.

REFERENCES

[1] Mohammad, A.S., A. Rattani, and R. Derahkshani. Eyeglasses detection based on learning and non-learning based classification schemes. in 2017 IEEE International Symposium on Technologies for Homeland Security (HST). 2017.

[2] Shao, L., R. Zhu, and Q. Zhao, Glasses Detection Using Convolutional Neural Networks, in Biometric Recognition: 11th Chinese Conference, CCBR 2016, Chengdu, China, October 14-16, 2016, Proceedings, Z. You, et al., Editors. 2016, Springer International Publishing: Cham. p. 711-719.

[3] Du, S., et al., Precise glasses detection algorithm for face with in-plane rotation. Multimedia Systems, 2017. 23(3): p. 293-302.

[4] Fernández, A., R. Casado, and R. Usamentiaga. A Real-Time Big Data Architecture for Glasses Detection Using Computer Vision Techniques. in 2015 3rd International Conference on Future Internet of Things and Cloud. 2015.

[5] 5. Liu, Z., et al. Deep learning face attributes in the wild. in Proceedings of the IEEE International Conference on Computer Vision. 2015.

[6] Sharif Razavian, A., et al. CNN features off-the-shelf: an astounding baseline for recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2014.

[7] Jing, Z. and R. Mariani. Glasses detection and extraction by deformable contour. in Pattern Recognition, 2000. Proceedings. 15th International Conference on. 2000. IEEE.

[8] Bo, W., A. Haizhou, and L. Ran. Glasses detection by boosting simple wavelet features. in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. 2004.

[9] Phillips, P.J., et al., The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing, 1998. 16(5): p. 295-306.

[10] Fernández, A., et al., Glasses detection on real images based on robust alignment. Machine Vision and Applications, 2015. 26(4): p. 519-531.

[11] Huang, G.B., et al., Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2007, Technical Report 07-49, University of Massachusetts, Amherst.

[12] Gao, W., et al., The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 2008. 38(1): p. 149-161.

[13] 13. Zheng, Y. Face detection and eyeglasses detection for thermal face recognition. 2012.

[14] Szegedy, C., et al. Going deeper with convolutions. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[15] Zhang, K., et al., Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016. 23(10): p. 1499-1503.

[16] Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[17] Shangfei, W., et al., A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference. Multimedia, IEEE Transactions on, 2010. 12(7): p. 682-691.

[18] Jia, Y., et al. Caffe: Convolutional architecture for fast feature embedding. in Proceedings of the 22nd ACM international conference on Multimedia. 2014. ACM.

[19] Kanwal, N., E. Bostanci, and A.F. Clark, Evaluation Method, Dataset Size or Dataset Content: How to Evaluate Algorithms for Image Matching? Journal of Mathematical Imaging and Vision, 2016. 55(3): p. 378-400.

978-1-5386-3007-5/17/$31.00 ©2017 IEEE