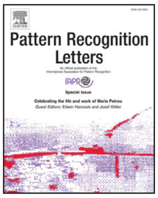
Pattern Recognition Letters 70 (2016) 59–65

Contents lists available at ScienceDirect

Pattern Recognition Letters

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



Minimalistic CNN-based ensemble model for gender prediction from face images[☆]

Grigory Antipov[a,b,*], Sid-Ahmed Berrani[a], Jean-Luc Dugelay[b]

[a] *Orange Labs - France Telecom, 4 rue du Clos Courtel, 35512 Cesson-Sévigné, France*
[b] *Eurecom, 450 route des Chappes, 06410 Biot, France*

ARTICLE INFO

Article history:
Received 4 August 2015
Available online 28 November 2015

Keywords:
Gender recognition from face images
Convolutional neural networks
Neural networks optimization

ABSTRACT

Despite being extensively studied in the literature, the problem of gender recognition from face images remains difficult when dealing with unconstrained images in a cross-dataset protocol. In this work, we propose a convolutional neural network ensemble model to improve the state-of-the-art accuracy of gender recognition from face images on one of the most challenging face image datasets today, LFW (Labeled Faces in the Wild). We find that convolutional neural networks need significantly less training data to obtain the state-of-the-art performance than previously proposed methods. Furthermore, our ensemble model is deliberately designed in a way that both its memory requirements and running time are minimized. This allows us to envision a potential usage of the constructed model in embedded devices or in a cloud platform for an intensive use on massive image databases.

1. Introduction

The human's gender plays a fundamental role in social interactions. Automatic gender classification has many important applications like intelligent user interface, visual surveillance, collecting demographic statistics for marketing, etc. Therefore, automatic gender recognition from face images has been extensively studied in computer vision. However, the difficulty of this problem largely depends on the application context and on the experimental protocol: a recognition model can be trained and tested on faces from the same dataset or from different datasets (i.e. cross-dataset experiment), images of input faces can be taken under controlled or uncontrolled conditions and finally faces can be aligned before gender prediction or not. The state-of-the-art performance in the most stringent conditions (i.e. cross-dataset, in uncontrolled environment and with no image preprocessing) reaches 96.86% of accuracy and was very recently obtained by Jia and Cristianini [11] using a huge private training dataset of 4,000,000 images.

Deep Convolutional Neural Networks (CNNs) [13] have recently become the golden standard for object recognition [12,25]. Today, CNNs are the primary choice for the large variety of computer vision tasks [8,27,30]. However, there are 2 problems which make the practical usage of CNNs difficult in some cases. The first problem is related to the big size of the training data which is often required to train them. Collecting large datasets of faces can be costly and can raise a number of privacy protection issues. That is why, successful face-related applications of CNNs are often trained on huge private datasets containing several millions of images (like in [27]) making the obtained results non-reproducible for the scientific community. The second problem lies in the domain of the computational and memory requirements of CNNs [7,9]. This problem often hinders importing CNNs onto embedded platforms like smartphones and tablets or their usage in cloud computations. For example, 16-layers CNN described in [25] has a weights file bigger than 500MB and requires about $3.1 \cdot 10^{10}$ floating point operations per image. Specifically, 90% of its weights is taken up by the fully-connected layers and more than 90% of its running time is taken by the convolutional layers [9]. It means that if we want to minimize both the running time and the required memory we have to minimize both fully-connected and convolutional layers.

In this work, we address the problem of gender recognition from face images taking into account the memory and the running time issues and by using a relatively small training dataset. In particular, we design a CNN-based ensemble model obtaining the state-of-the-art performance on gender recognition from face images in the most stringent conditions. We use a publicly available dataset of face images to train our CNN-model obtaining the highest recognition accuracy with about 10 times less training data than the state-of-the-art authors [11]. Our model is also minimized both in terms of the running time and the memory requirements making its usage possible even on devices with a

[☆] This paper has been recommended for acceptance by R. Davies.

[*] Corresponding author at: Orange Labs - France Telecom, 4 rue du Clos Courtel, 35512 Cesson-Sévigné, France. Tel.: +33 2 99 12 41 88; fax: +33 2 99 12 40 98.

E-mail address: grigory.antipov@orange.com (G. Antipov).

http://dx.doi.org/10.1016/j.patrec.2015.11.011
0167-8655/

limited memory and without dedicated graphical processors for computations.

The rest of the paper is organized as follows: the overview of the related literature is done in Section 2; the datasets used for training and test are presented in Section 3; the Starting CNN and the methodology to progressively minimize it are proposed in Section 4; the procedure of minimization of the Starting CNN is described in Section 5; the classification results are analyzed in Section 6; and the conclusions are summarized in Section 7.

2. Related work

In this section, we make on overview of existing works on gender recognition from face images.

Early works on gender recognition from face images focused on the case of frontal faces in a controlled laboratory environment. In the beginning of the 90s, many authors tried neural networks to deal with this problem. For example, Golomb et al. [6] trained a 2-layers fully-connected neural network and achieved 91.90% accuracy on a tiny test set of 90 images. The benchmark dataset of frontal faces in a controlled environment is FERET [20]. With the emergence of SVM, Moghaddam and Yang [18] used this classifier with an RBF kernel on raw pixels and obtained 96.62% accuracy on FERET (though having the same persons presented both in training and test sets). Rather than using SVM, Baluja and Rowley [2] used AdaBoost on raw pixels and obtained 96.40% on FERET without mixing people in training and test sets. Li et al. [15] combined facial information with clothing and hair components obtaining 95.10% accuracy on the FERET dataset. Ullah et al. [29] used the Webers Local texture Descriptor to reach almost perfect performance of 99.08% on FERET. This result suggests that the FERET benchmark is saturated and not enough challenging for modern methods.

As a result, the majority of contemporary works deals with the problem of gender recognition from face images in an uncontrolled environment. The Labeled Faces in the Wild (LFW) dataset [10] is the most frequently used one in this case. Different works on gender recognition in an uncontrolled environment are compared in Table 1. Shan [23] employed Local Binary Patterns (LBP) features with an AdaBoost classifier to obtain 94.81% on LFW. Shih [24] used the Active Appearance Model (AAM) in order to align face images and to model them using small patches around the detected landmarks. The Bayesian framework was employed as a classifier. The resulting model obtained 86.50% classification accuracy on the combination of the color FERET and LFW datasets. Tapia and Perez [28] fused LBP features with different radii and spatial scales and used an SVM classifier above. The authors performed 2 experiments: in the first one, they trained and tested their models on different subsets of LFW, while in the second one, the training was done on a separate dataset. Results of these 2 experiments (95.60% and 98.01%) differ quite significantly from each other proving that the cross-database protocol is more challenging. Bekios-Calfa et al. [4] showed that it may be advantageous to predict the person's gender simultaneously with the person's age and pose in the photo. They got 79.11% gender recognition accuracy training their model on the GROUPS dataset and testing on the LFW dataset. The most recent attempt to employ CNNs for gender recognition from face images was done by Levi and Hassner [14]. Authors trained a CNN on the newly created Adience dataset. They obtained a relatively modest accuracy of 86.80% mainly because of the low quality of images in Adience. Finally, the most recent result on the LFW dataset under the cross-database protocol was obtained by Jia and Cristianini [11]. The authors used a huge private dataset of 4,000,000 images to train a C-Pegasos classifier (a variation of SVM) using LBP fetaures. They obtained a state-of-the-art accuracy of 96.86% on LFW referring their success mainly to the size of the training dataset.

In this work, we use the result of Jia and Cristianini as a baseline for comparison with our models.

It should be mentioned that face images are by far not the only possible modality to predict a person's gender. There are works on gender predictions from gait [5,16], speech [17], images of silhouettes [1] and even web forum messages [34]. However, in this work, we focus only on the gender prediction from face images and therefore do not consider other modalities.

3. Datasets

In this section, we present face datasets which have been used in our experiments.

We have used 2 publicly available face datasets: CASIA WebFace and Labeled Faces in the Wild (LFW). The first one is used for training and validation whereas the second one is used only for testing. While collecting the CASIA WebFace dataset, its authors made sure that there are no subject intersections between CASIA WebFace and LFW [32].

3.1. CASIA WebFace dataset

CASIA WebFace dataset was collected for the face recognition purposes by Yi et al. [32]. The dataset contains photos of actors and actresses born between 1940 and 2014 from the IMDb website.[1] Images of the CASIA WebFace dataset include random variations of poses, illuminations, facial expressions and image resolutions. In total, there are 494,414 face images of 10,575 subjects. To the best of our knowledge, CASIA WebFace is the biggest publicly available face dataset today, and that is why we have used it to train CNNs in this work.

Authors of CASIA WebFace provide names of 10,575 subjects but not their genders. We have annotated genders using the metadata provided by IMDb and also by manual annotation.

3.2. LFW dataset

Being collected by Huang et al. [10], the LFW dataset has become a benchmark for face gender recognition in an unconstrained environment. It consists of 13,233 face images of 5749 celebrities. Contrary to CASIA WebFace, LFW does not only contain photos of actors and actresses but it also contains photos of politicians, sportsmen and sportswomen.[2]

3.3. Data preprocessing

Images of both CASIA WebFace and LFW are face-centered and have an initial resolution of 250 × 250 pixels. The two datasets have been processed in the same way: the faces are firstly extracted with the Viola–Jones face detector [31], and then they are rescaled to a certain square size (the particular size depends on the input dimensions of a CNN). This process is illustrated in Fig. 1. In case if several faces are found in an image, only the biggest one is taken; if no faces are found in an image, the image is ignored. After face extraction, we have obtained 452,042 face images from the CASIA WebFace dataset. These images have been split into training and validation sets in the proportion of 95% and 5%, respectively. We have ensured that there are no subject intersections between training and validation sets. In order to be able to fairly compare our results with the current state-of-the-art in gender recognition on LFW, we have used exactly the same test set of 10,147 face images as the authors of the current best result on LFW [11]. Following their work, we have not performed any sort of alignment to the test images prior to gender classification. More details on the data split into training, validation and test sets are given in Table 2.

[1] http://www.imdb.com/ : Intenet Movie Database (IMDb) is an online database of information related to films, television programs and video games.

[2] Gender annotations for the LFW dataset are available at http://face.cs.kit.edu/431.php .