Pattern Recognition Letters 58 (2015) 35–41

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Learning to classify gender from four million images ☆

Sen Jia*, Nello Cristianini

Intelligent Systems Laboratory, University of Bristol, Merchant Venturers Building, Woodland Rd, Bristol BS8 1UB, UK

ARTICLE INFO

Article history:
Received 18 September 2014
Available online 26 February 2015

Keywords:
Big data
Gender classification
On-line learning

ABSTRACT

The application of learning algorithms to big datasets has been identified for a long time as an effective way to attack important tasks in pattern recognition, but the generation of large annotated datasets has a significant cost. We present a simple and effective method to generate a classifier of face images, by training a linear classification algorithm on a massive dataset entirely assembled and labelled by automated means. In doing so, we perform the largest experiment on face gender recognition so far published, reporting the highest performance yet. Four million images and more than 60,000 features are used to train online classifiers. By using an ensemble of linear classifiers, we achieve an accuracy of 96.86% on the most challenging public database, labelled faces in the wild (LFW), 2.05% higher than the previous best result on the same dataset (Shan, 2012). This result is relevant both for the machine learning community, addressing the role of large datasets, and the computer vision community, providing a way to make high quality face gender classifiers. Furthermore, we propose a general way to generate and exploit massive data without human annotation. Finally, we demonstrate a simple and effective adaptation of the Pegasos that makes it more robust.

1. Introduction

The task of face classification has been extensively studied in computer vision. As many other similar tasks, its difficulty varies greatly with the particular conditions in which the images are obtained. In this domain, a standard dataset has been used as a benchmarking tool for this task in the past few years, LFW [2], so for comparability we measure our performance on that dataset, using more stringent experimental settings than those reported in Ref. [1]. Importantly, we test on the LFW for comparability, but we do not rely on it for training. Instead, four million face images collected from the web and high dimensional features are used in this paper.

Halevy et al. [3] identify the enabling factor behind much recent success in machine intelligence as the application of simple machine learning methods to very large training sets. The expression "the unreasonable effectiveness of data" is used in that article to express the observation that "simple models based on more data trump more elaborate models based on less data". This obviously creates the problem of gathering, curating and annotating very large training sets, causing the authors to remark "the first lesson of web-scale learning is to use available data rather than hoping for annotated data that is not available", and to recommend the use of data already "available in the wild" (that is data not created with research questions in mind but as the by-product of other activities). This is the approach we pursued to train the face gender classifiers described in this study: using very large amounts of training images gathered from the web and labelled automatically, and simple large margin linear classifiers. This study is by far the largest study yet published on face gender classification, training a classifier on four million images collected from the web, and described with over 60,000 features, we named it 4 million weakly labelled faces in the wild (4MWLFW). We use large margin linear classifiers training in on-line fashion, either alone or in ensemble. The size of the training set makes it infeasible to compare with SVM solutions. We report the highest performance yet observed on this task, as measured on the standard LFW benchmark: an accuracy of 96.86%, compared to the previous best result of 94.81%.

Achieving such a high performance with a simple linear classifier seems to lend support to the "unreasonable effectiveness of data" conjecture of Halevy et al. [3]. Our experimental findings are relevant both for the computer vision and for the machine learning communities, relating to the current interest in the use of deep classifiers, to the discussion about feature descriptors for face images [4], and generally to the potential of big-data to enable various machine intelligence tasks.

The rest of the paper is organised as follows: In Section 2, we describe the data collection and curation procedure. In Section 3, we describe the feature extraction and classification algorithm. In Section 4, we describe the experimental results. In Section 5, we discuss the implications of this result and the relations with other research.

☆ This paper has been recommended for acceptance by Jie Zou.

* Corresponding author. Tel.: +44 7851463027.
E-mail address: jason.jia@bristol.ac.uk (S. Jia).

http://dx.doi.org/10.1016/j.patrec.2015.02.006
0167-8655/

Table 1
Related work and their specifications.

Methods	Accuracy (%)	Real life	Cross database	Automatic	Sample size	Feature and classifier
Our method	96.86	✔	✔	✔	4 million	Multi-scale LBP + C-Pegasos
[5]	75.10	✔	✔	✔	17, 814	LBP + linear SVM
[6]	89.77	✔	✔	✔	14, 760	LBP + PCA + SVM
[7]	88	✔	✔	✔	13, 600	Pixels + LUT AdaBoost
[1]	94.81	✔	✔	✗	7443	Boosted LBP + SVM
[8]	79	✔	✗	✗	About 3500	Haar-like + AdaBoost
[9]	86.54	✗	✔	✔	411	Pixels + SVM
[10]	91.3	✗	✔	✔	10, 669	Pixels + RBF SVM
[11]	92	✗	✔	✔	11, 500	Haar-like + real AdaBoost
[12]	94.3	✗	✔	✔	2409	Pixel comparisons + AdaBoost
[13]	96.62	✗	✗	✔	1755	Pixels + SVM

1.1. Related works in face gender classification

The performance of face gender classifiers depends greatly on the experimental conditions in which they are tested. Earlier studies were conducted under very constrained and stable settings (e.g, constant position, expression and lighting) and obtained high performance. But as soon as more realistic settings were introduced, it was clear that the task was very dependent on them. The current state of the art requires that classifiers are tested on real-world images, with variations in lighting, face expressions and backgrounds. To make the test more realistic, it is also necessary to ensure that images of the same person are not included both in the training and in the testing set, and finally that no human alignment, rectification or other processing of examples is performed. Our study will follow the most stringent experimental set-up (Table 1).

Previous work addressing the task of face gender classification in images started out using constrained face database, the FERET [14]. The best result on this database was 96.62%, obtained by Moghaddam and Yang [13] by combining pixel intensities with a radial basis function (RBF) kernel support vector machine (SVM). However, a few years later Baluja and Rowley [12] showed that results are biased when the same person appears in both the training and test sets. After manually separating the FERET database so no person appeared in both sets, they achieved an accuracy of 94.3% using efficient pixel comparisons and AdaBoost.

In most cases, we do not wish to manually align faces or remove false detected images due to the prohibitive cost, favouring automated approaches. Wu et al. [7] combined web face images with the FERET database to test an automatic gender classifier, achieving 88% without the need for manual intervention. Similarly, Yang et al. [11] present an automatic gender classification system, obtaining 92% on the FERET database using real AdaBoost. More recently, Makinen and Raisamo [9] performed an experiment based on various combinations of alignment locations, with the best result achieving 86.54% on the FERET database.

An early attempt using face images retrieved from the web demonstrated the increased difficult in classifying gender from real life images, with Shakhnarovich et al. [8] achieving 79% accuracy using their more realistic web image database after removing faces which are more than 30° from frontal. More recent studies using web images for training [10] achieved 91.3% on the FERET database.

In 2007, Huang et al. [2] built a public database for face-based classification tasks, named labelled faces in the wild (LFW), to enable the testing of algorithms with a standard, realistic benchmark. Using the MORPH database [15], Ramn-Balmaseda et al. [5] achieved an accuracy of 75.10% on LFW, while Dago-Casas et al. [6] achieved 89.77% training on the web-based database [16], both using local binary patterns (LBP) for their features. The highest result on LFW is by Shan [1] using boosted LBP features to train an SVM, achieving an accuracy of 94.81% using a commercially aligned version of LFW [17].

2. Datasets and description

2.1. Data acquisition

The task of collecting, curating and annotating large amounts of data is the main bottleneck in "web-scale learning", and the development of automated means to leverage data "existing in the wild" is of central importance to that approach. In this section we describe how we collected and curated the training data, and which data we choose as a test set.

Training set (4MWLFW): We have obtained images by querying search engines with gender specific queries, such as "John" for male faces and "Mary" for female faces. These gender specific queries were obtained by making use of a long list of names and relative gender labels provided by Internet Movie Database (IMDB).[1] These are specific names such as "Tom Cruise" in the male list, and "Nicole Kidman" in the female list. Each query is used to retrieve a set of images from the search engine, all of which inherit the gender label of the query itself, and is then fed through a software pipeline.

Assessing the extent to which a given query is gender-specific is an interesting problem to automate, one for which many heuristics can be readily proposed, but for the sake of this article we will make use of a pre-fixed set of queries with high-confidence gender labels.

As described in Ref. [2], we begin by detecting the position of the faces and facial landmarks in these images, using the Viola–Jones detection [18], implemented within OpenCV.

The images of faces are then rescaled to a standard size of 90 × 120 pixels and turned into grey-scale images. In this way we have a standardised image of a face, and a tentative gender label coming from the query used to retrieve it. The images so obtained were four million, and the gender balance was approximately of 50% male and 50% female labels.

Manual cleaning, removal of spurious images, or manual relabelling, could be used here to improve quality, but we do not make use of these steps, in order to obtain a fully automated extraction pipeline that can be used to generate vast datasets for training.

Test set: As a test set we used the standard LFW dataset [2], in order to ensure comparability with previous work. Using LFW as a test set means that we operated in the difficult situation of training on a dataset and testing our performance on a different one, in this way emulating real operating conditions of a gender classifier. Both our training and our testing images were taken from the web and therefore in uncontrolled conditions. LFW can be considered as a

[1] ftp://ftp.fu-berlin.de/pub/misc/movies/database/.