Neurocomputing 174 (2016) 1077–1086

Contents lists available at ScienceDirect

Neurocomputing

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

A new descriptor of gradients Self-Similarity for smile detection in unconstrained scenarios ☆

Yuan Gao, Hong Liu *, Pingping Wu, Can Wang

Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 11 July 2015
Received in revised form
11 September 2015
Accepted 8 October 2015
Communicated by Huaping Liu
Available online 17 October 2015

Keywords:
Smile detection
Histogram of Oriented Gradients
Self-Similarity of Gradients
AdaBoost
Support Vector Machine
Extreme Learning Machines

ABSTRACT

Smile detection is a sub-problem of facial expression recognition field, which has attracted more and more interests from researchers because of its wide application market. As for smile detection problem itself, the 'wild' unconstrained scenario is more challenging than the laboratory constrained scenario. Therefore, in this paper, we mainly focus on solving smile detection problem in unconstrained scenarios. To this end, a new descriptor, Self-Similarity of Gradients (GSS), is proposed. Inspired by Self-Similarity on Color channels (CSS) feature in pedestrian detection area, GSS can effectively describe the similarities in a HOG feature map, while these similarities are useful and helpful for constructing a high-performance practical smile detector. Moreover, since a smile detector using multiple features and multiple classifiers simultaneously shows superior performance, they are also adopted by us. Finally, experimental results indicate that the combined features (HOG31+GSS+Raw pixel) using AdaBoost with linear Extreme Learning Machines (ELM) achieve improved performance over the state-of-the-arts on the real-world smile dataset (GENKI-4K).

1. Introduction

For human beings, smile is one of the most common expressions. Smile detection has a lot of underlying applications, such as smile shutter function in digital cameras, expression understanding in human-robot interaction, user expression feedback and statistics in Kinect-type interaction games.

Typically, smile detection in unconstrained scenarios is a challenging problem, which is because that imaging conditions of real-world scenarios, e.g., illumination and occlusion, are much more complex than those in laboratory environments. A well-trained model on databases of laboratory environment surprisingly behaves badly on databases in the 'wild' condition, while for the model trained on real-world databases, the conclusion is reversed [1]. Therefore, in this paper, we study smile detection problem in unconstrained scenarios. A real-world database, GENKI-4K, the only publicly available dataset in unconstrained scenarios for smile detection in research area, is selected to perform our experiments.

For smile detection in unconstrained scenarios, feature representation is the key step. A lot of traditional feature representation methods, such as PCA [2], LDA [3], Gabor [4], Haar [5], LBP [6], LPQ [7] and HOG [8], have been utilized to solve this problem. Recently, a variant HOG [9] is proposed and has become a promising feature for many computer vision problems. To the best of our knowledge, we are the first to use it in smile detection tasks and it achieves better performance compared with other baseline features.

Inspired by the big success of Self-Similarity on Color Channels (CSS) [10] in pedestrian detection area, in this paper, we also find some similarities in a HOG feature map after visualizing high-dimensional HOG feature of face images. As has been shown in [10], encoded similarities are an important kind of supplement feature to improve a pedestrian detector's performance. Therefore, we propose to use Self-Similarity of Gradients (GSS) feature to describe and encode similarities in face images. Apart from this, in the face registration procedure, eyes-and-mouth-based alignment is proven to be more effective than eyes-based alignment for a smile detector. Then in the step of classification, the smile detector using classifier combination shows better performance than those only using one type of classification method. Finally, the best smile recognition rate in unconstrained scenarios is achieved by using feature combination (HOG31+GSS+Raw pixel) and classifier combination (AdaBoost+Linear ELM) strategies simultaneously.

☆This work is supported by National Natural Science Foundation of China (NSFC, No. 61340046), National High Technology Research and Development Program of China (863 Program, No. 2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, JCYJ20130331144716089), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011).

* Corresponding author.
E-mail addresses: ygao@sz.pku.edu.cn (Y. Gao), hongliu@pku.edu.cn (H. Liu), wupingping@pku.edu.cn (P. Wu), canwang@pku.edu.cn (C. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.10.022
0925-2312/

The remainder of this paper is organized as follows. We review related works in Section 2. Subsequently, Section 3 involves three important steps in smile detection, which are face registration, feature representation and classification. Experiments and analysis are described in Section 4. And conclusions are indicated in Section 5.

2. Related work

Although there do not exist many literatures dedicated to smile detection, it is still an important part of automatic facial expression analysis, which is a mature research field in computer vision area. For automatic facial expression analysis, the standard processing pipeline is composed of four steps, including face detection, face registration, feature extraction and classification [11]. Among them, the last two procedures are certain research hotspots. Specifically, in feature extraction, geometric (or shape)-based features [12,13] and appearance-based features [14,15] are commonly extracted. And in classification, three different types of binary classifiers are usually employed, which are Artificial Neural Networks (ANN), ensemble learning techniques and Support Vector Machines (SVM).

Regarding specific smile detection problem, most existing research works focus on the improvements in both of these two procedures. Shinohara et al. got effective features from Higher-order Local Auto-Correlation (HLAC) features using Fisher Weight Map (FWM) and achieved better performance compared with Fisherfaces method and HLAC-features-based method for smile detection on their own database of only four people [16]. Bai et al. extracted Pyramid Histogram of Oriented Gradients (PHOG) features from the region of mouth and achieved as high a smile detection rate as Gabor features did on Cohn-Kanade AU-Coded Facial Expression Database [17]. Nevertheless, both of them executed experiments on databases under constrained laboratory environments. A comprehensive work for smile detection in unconstrained or wild scenarios was proposed by Whitehill et al., and it was also the basis for smile detection function of modern digital cameras [1]. At the same time, a new dataset with contents from the web, namely GENKI, was made public by them for smile detection research in the real-world condition. On this dataset, Shan proposed a novel smile detection approach by simply comparing the intensities of a few pixels in a face image and achieved better performance than Gabor+SVM [18,19]. And Zhang et al. found that only using Mouth Feature (MF) could achieve comparable smile detection performance with the whole face image using intensity difference, Maximum Feature Difference (MFD) and AdaBoost algorithms but significantly reduced the computing and memory consumptions simultaneously [20]. More recently, An et al. showed that with the same features extracted from faces, Extreme Learning Machines (ELM) outperformed Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) both on GENKI-4K and their own collected MIX databases [21].

3. Technical approach

The whole method involves three steps, which are face registration, feature representation, and classification. Details of these three steps are introduced as below:

3.1. Face registration

Image registration is shown to be one of the vital procedures of developing a high-performance smile detector [1]. A preliminary for image registration is the detection of important facial landmarks. For example, locations of two eyes have to be found in [19] and [1]. As for facial landmark detection, numerous methods have been proposed [22–25]. Recent research work [26] has shown its effectiveness and efficiency in facial landmark detection problem, and it can even handle well with partial or uncertain labels. Obviously, [22] and [26] could be directly used in a practical real-time smile detector. Nevertheless, an accurate face landmark detector depends highly on an accurate face detector for initialization, and in this paper, we mainly care feature-related effect for a smile detector. Therefore, the manual manner is finally chosen.

After this, it is very important to decide which facial landmarks points need to be labeled. Typically, labeling the centers of eyes is a common way. But when observing some result examples in this way, e.g., face images in the first and third rows of Fig. 1, it can be clearly found that some parts of faces have been truncated, especially the mouth part. To be more precise, the eyes-based face alignment method leads to the discrepancy of mouth positions. As we all known, image information of mouths must be significant for a image-based smile detector. Therefore, mouths also need to be aligned as eyes. Based on the above observation and analysis, we propose to utilize an eyes-and-mouth-based face alignment manner, details of which have been shown in Fig. 2. Some result examples using this method are illustrated in the second and fourth rows of Fig. 1. Compared with results of eyes-based alignment, the details of mouths are entirely reserved, which lays a solid basis for the subsequent feature extraction process.

Finally, no matter eyes-based or eyes-and-mouth-based face alignment method, affine transform matrixes could be easily computed using the positions of labeled facial landmark points. Specifically, affine transform is composed of rotating, cropping and scaling. And 48×48 pixels are the output resolution for all images.

3.2. Feature representation

Since GSS feature absolutely relies on the pre-calculation of HOG feature, both of HOG and GSS features are described sequently in this subsection. Besides, HOG visualization is important for constructing a GSS descriptor, so it will also be introduced in detail.

3.2.1. HOG36

Histogram of Oriented Gradients (HOG) are originally proposed by Dalal and Triggs for pedestrian detection problem [8]. For a gray-scale input image ($w \times h$ resolution), the gradients of it could be computed using $[-1, 0, +1]^T$ and $[-1, 0, +1]$ filters. Then the gradient orientation and magnitude of pixel (x, y) could be represented as $\theta(x, y)$ and $r(x, y)$. Afterwards, a new matrix B_1 indicating contrast insensitive is shown as follows:

$$B_1(x, y) = round\left(\frac{p\theta(x, y)}{\pi}\right) mod\ p \tag{1}$$

B_1 has the same size as the source input image. Here, p stands for the number of orientation bins. After this, the gradients image could be indicated as a $w \times h \times p$ sparse feature map F:

$$F(x, y, z) = \begin{cases} r(x, y) & \text{if } z = B_1(x, y) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Subsequently, the feature map F need to be transformed into a cell-based feature map C, and any cell is in the size of $c \times c$. So in $C(i, j, k)$, i meets the $0 \le i \le \lfloor (w-1)/c \rfloor$ condition, j meets the $0 \le j \le \lfloor (h-1)/c \rfloor$ condition, and k meets the $0 \le k \le p-1$ condition. Besides, $C(i, j)$ is actually the sum of all the p-dimensional items of F in the corresponding (i, j) cell. In the normalization step, every feature vector $C(i, j)$ has four different normalization factors which