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

Fig. 1. Face registration results using eyes-based and eyes-and-mouth-based face alignment methods. In particular, images in row one and row three are results processed by the eyes-based face alignment method. And images in row two and row four are results processed by the eyes-and-mouth-based face alignment method.

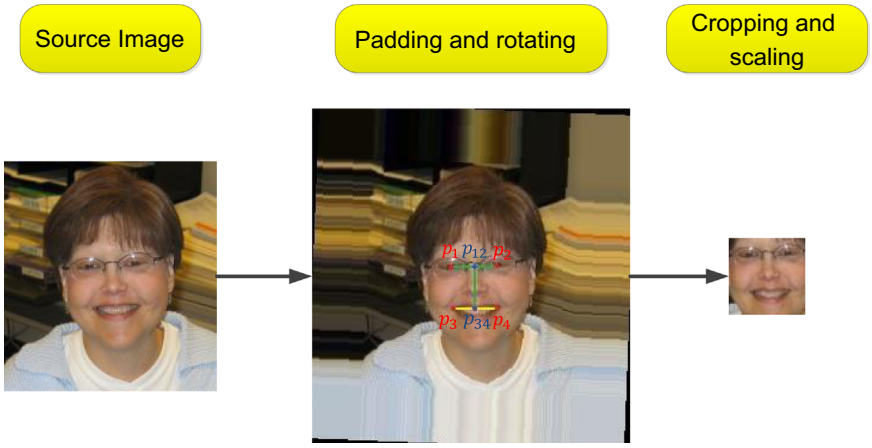


Fig. 2. The procedure of affine transform based on the detected facial landmarks.

could be written as

$$N_{\delta,\gamma} = \left(\| C(i,j) \|^2 + \| C(i+\delta,j) \|^2 + \| C(i,j+\gamma) \|^2 + \| C(i+\delta,j+\gamma \|^2 \right)^{1/2} \tag{3}$$

here, $\delta,\gamma \in \{-1,1\}$. Finally, each cell-based feature vector $C(i,j)$ generates a cell-base HOG matrix $H_1(x,y)$:

$$H_1(i,j) = \begin{bmatrix} T_\alpha(C(i,j)^T / N_{-1,-1}(i,j)) \\ T_\alpha(C(i,j)^T / N_{+1,-1}(i,j)) \\ T_\alpha(C(i,j)^T / N_{+1,+1}(i,j)) \\ T_\alpha(C(i,j)^T / N_{-1,+1}(i,j)) \end{bmatrix} \tag{4}$$

Basically, $T_\alpha(v)$ is a truncation function which means that every element in $T_\alpha(v)$ is no larger than α.

Typically, $p=9, c=8, \alpha=0.2$ are standard settings in a lot of works. Therefore, $H_1(x,y)$ is commonly a 4×9 matrix, which could be transformed into a 36-dimensional vector and it is one item in the final cell-based HOG feature map. In this paper, we call this type of HOG as HOG36.

3.2.2. HOG31

A variant type of HOG36 feature has shown its effectiveness in object detection tasks [9]. Come back to the B_1 matrix which indicates contrast insensitive, how about the contrast sensitive condition? We use B_2 stand for it:

$$B_2(x,y) = round\left(\frac{q\theta(x,y)}{2\pi} \right) \ mod \ q \tag{5}$$

For the contrast sensitive condition, q is generally two times of p. So q is equal to 18. After the same operations as above on B_2, a new cell-based HOG matrix $H_2(x,y)$ could be got and it is a 4×18 matrix. Then, let $H(x,y)$ be the combination of $H_1(x,y)$ and $H_2(x,y)$:

$$H(x,y) = [H_1(x,y), H_2(x,y)] \tag{6}$$

Consequently, $H(x,y)$ is a 4×27 matrix. By computing the sum of the elements in each row and each column of $H(x,y)$, a 31-dimensional vector could be obtained. And it is finally one member in the cell-based HOG feature map. As a result, we use HOG31 to stand for this variant HOG.

3.2.3. HOG visualization

Because of having the ability of describing shape information of images effectively, HOG feature is one of the most important features in object detection area. In order to observe what is happening in high-dimensional HOG world, it is meaningful to visualize HOG features, which helps researchers to obtain a better understanding of the behaviors of these HOG-based object detectors.

Here, two different types of HOG visualization methods implemented on face images are illustrated and compared in Fig. 3. In particular, (a) and (b) in the first column stand for the "mean" faces of non-smile and smile images in GENKI-4K. Traditional HOG glyphs of (a) and (b) are shown in (c) and (d) respectively. In the third column, (e) and (f) are the results of using the HOG visualizing method of Vondrick et al. [27] corresponding to (a) and (b) severally.

When observing the visualized gray-scale images (c)–(f) in Fig. 3, we find that the differences between non-smile and smile "mean" faces are mainly distributed in the mouth, cheeks and eyes regions. In addition, the mouth region and eyes region in the non-smile "mean" are very similar. However, in the smile "mean" face, the observation result is different. It is quite obvious that the center of the mouth is different from the other parts of the mouth but is more like cheeks.

Subsequently, to quantify the detailed similarities in a HOG feature map, we calculate euclidean distance between the cell-based seed feature vector and other cell-based feature vectors. One example result is shown in Fig. 4, in which the cell-based seed is located in the fifth row and fourth column of cell-based HOG feature maps, and it exactly corresponds to the right side of the mouth. The biggest difference between these two quantified pictures is that, in the non-smile "mean" face image, the right side of the mouth is not very similar to the left side of the mouth, while in the smile "mean" face image, these two sides are similar. Intuitively, utilizing these pairwise statistics of localized gradient distributions may be beneficial to construct a smile detector with good performance.

To this end, we encode self-similarities between cell-based HOG feature vectors and name this kind of feature as Self-Similarity of Gradients, the details of which will be described in the end of this subsection.

3.2.4. Self-Similarity of Gradients

The overall framework of GSS feature extraction process is presented in Fig. 5. The input image is a well aligned picture using the eyes-and-mouth-based face alignment method as described in the previous subsection. Then, on this gray-scale input image, a cell-based HOG31 feature map ($m \times m \times 31$) can be computed easily. Subsequently, this cell-based HOG feature map is split into several block-based HOG feature maps. And a block-based HOG feature map has $n \times n$ cells with 31 dimensions in depth. Apparently, the cell-based HOG feature vectors in a block-based feature map could be compared and the comparison results exactly stand for the similarities of gradients. In each block-based feature map, there are $B_{compare}$ feature vector comparisons:

$$B_{compare} = C^2_{(n\times n)} = \frac{n^2 \times (n^2-1)}{2} \tag{7}$$

Besides, the cell-base stride between two nearest-neighboring blocks is denoted as k. So there are B_{num} blocks in the HOG feature map:

$$B_{num} = \left(\left\lfloor \frac{m-n}{k} \right\rfloor + 1\right)^2 \tag{8}$$

Finally, the output GSS feature is a $(B_{num} \times B_{compare})$-dimensional vector. With regard to the cell-based feature vector comparison functions, several methods have been tried. For example,

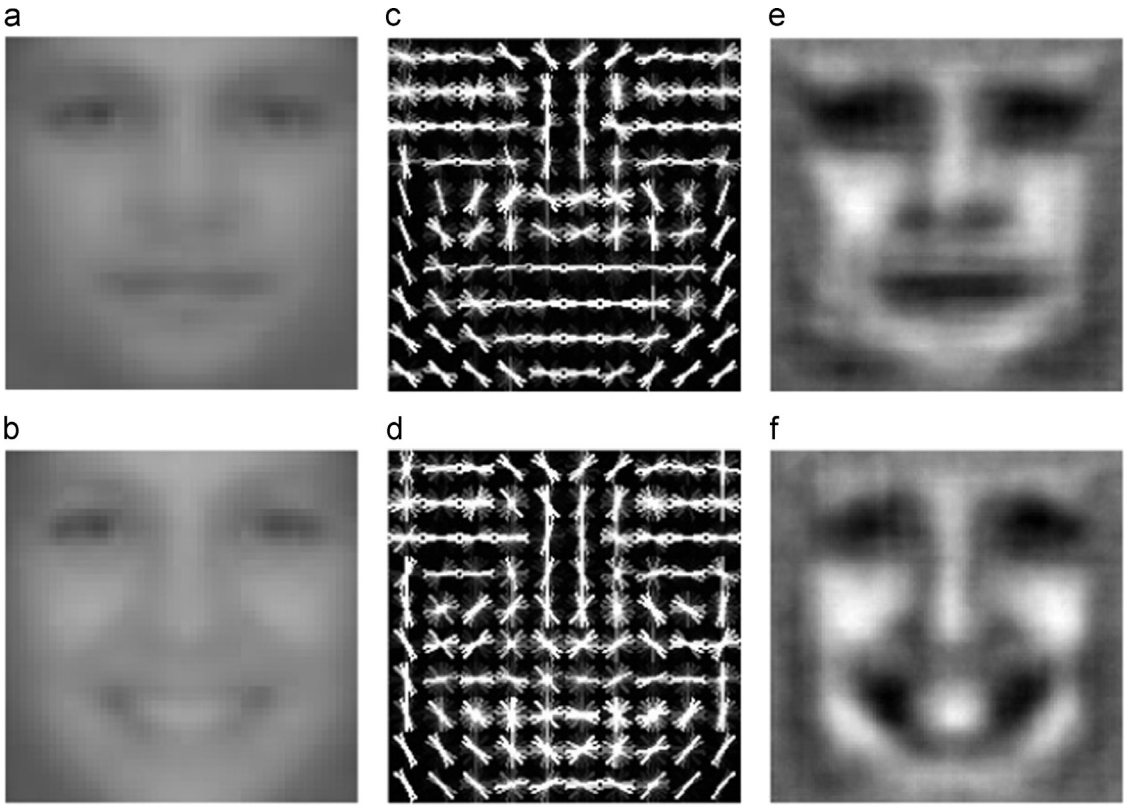


Fig. 3. Two different HOG visualization methods on "mean" faces of smile and non-smile in GENKI-4K.

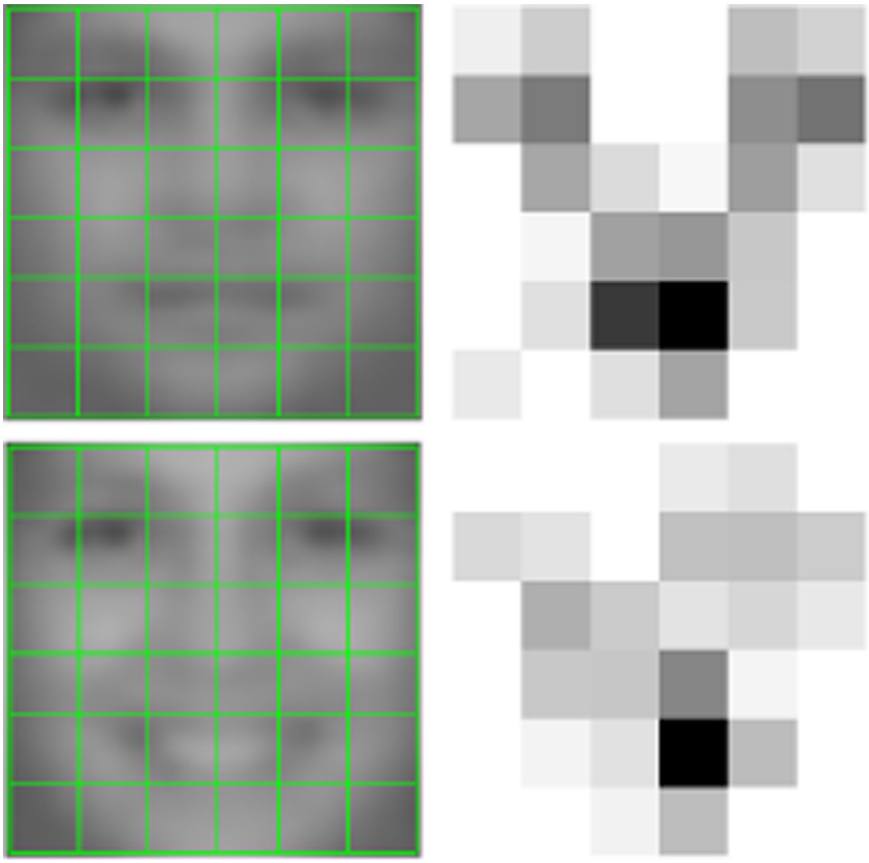


Fig. 4. The left column corresponds two "mean" faces of non-smile and smile in GENKI-4K. The right column corresponds to the similarity intensities between the seed cell (Row 5 and Column 4) and other cells in cell-based HOG feature maps of these two "mean" face images. The gray-scale visualized images actually show similarity intensities. The higher gray-scale intensity means the higher similarity.

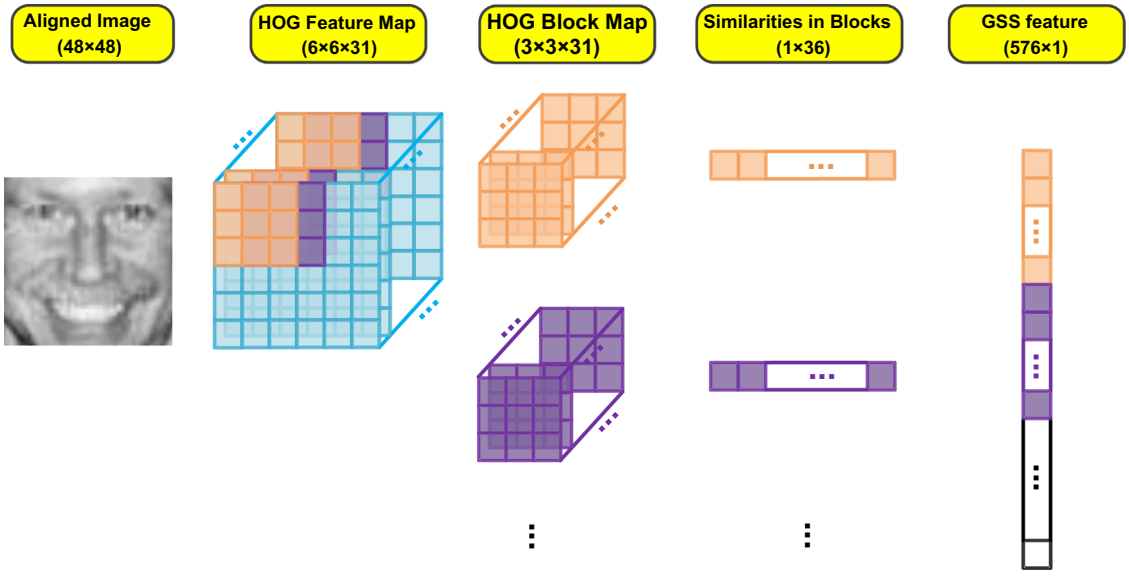


Fig. 5. Framework of extracting GSS feature.

Cityblock, Chebychev, Cosine, Correlation, Spearman, Euclidean, and Square Euclidean. And the Square Euclidean method is selected because of achieving the best performance for smile detection task. Details will be introduced in the experiment and analysis part.

3.3. *Classifiers*

Three different types of popular machine learning algorithms have been utilized in our work, which will be introduced briefly one by one.

3.3.1. Adaptive boosting

Adaptive boosting algorithm is originally proposed by Freund and Schapire [28]. The key idea is, in each iteration, through adjusting the weights of training samples, the most discriminative feature (weak learner) could be selected. Each weak learner is composed of a "stump" or a CART tree. It is worthy being noticed that AdaBoost has several variations, which are mainly different in the algorithm designment in the training step. In them, Gentle AdaBoost [29] is selected in our experiments for that it has been shown to be the most practically efficient boosting algorithm in some applications [30].

3.3.2. Support Vector Machine

Support Vector Machine is initially proposed by Cortes and Vapnik in 1995 with the soft margin idea [31]. The key to SVM in binary classification is to find several support vectors and, at the same time, maximize the geometric margin between positive training samples and negative training ones. Here, a linear SVM is selected in consideration of its good performance, simplicity and, last but not least, the speed, which could be exhibited using the this formulation:

$$f(x) = \mathrm{sgn}(w^T x + b) \tag{9}$$

Obviously, the complexity of this function is linear with the size of the testing vector x.

3.3.3. Extreme Learning Machines

Extreme Learning Machines (ELM) are firstly designed for generalized single-hidden layer feedforward networks (SLFNs) [32]. Since the excellent classification and regression performances comparable to SVM, ELM has attracted more and more attentions of researchers in the computer vision area. More recently, An et al. show that ELM performs better than other classifiers (e.g. SVM and LDA) for a smile detector [21]. So we also pay attention to this extremely outstanding classifier. But different to their work, we care about ELM in the kernel case, which is more promising in experiments compared with the early version they use [33].

The kernel version of ELM can be written as

$$f(x) = h(x)H^T \left(\frac{I}{C} + HH^T\right)^{-1} T \tag{10}$$

Here, x stands for an input feature vector. $h(x)$ is a function transforming x into a higher dimensional vector and users do not need to know how it works. Subsequently, H is a vector which has N values and $H_i = h(x_i)$. N is the number of training samples. Besides, I is a $N \times N$ identity matrix. C is the only parameter users have to set. Finally, T is a $N \times 2$ matrix indicating true labels for binary classification.

So when multiplying two different $h(x)$ together, there exists the same trick as SVM doing. Kernel function is this exact trick. Let Ω be the equal of HH^T, where

$$\Omega_{i,j} = h(x_i) \cdot h(x_j) = K(x_i, x_j) \tag{11}$$

Back to Eq. (10), a new one could be written as

$$f(x) = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_N) \end{bmatrix}^T \left(\frac{I}{C} + \Omega\right)^{-1} T \tag{12}$$

In the training procedure, $(I/C + \Omega)^{-1} T$ could be easily got utilizing N training sample vectors. After this, any input test vector's label could be instantly predicted using formulation (12). The position of the maximal value in $f(x)$ corresponds to which class the test vector will be in. For the sake of building a real-time smile detector, we do not use sophisticated kernels like kernel-SVM. Instead, a linear one is chosen by us and it could be written as

$$K(x_i, x_j) = x_i^T x_j \tag{13}$$

3.3.4. Classifier combination

Apart from using these three algorithms individually, the combination of AdaBoost and another classifier is also an effective and efficient way to improve classification performance. In this case, AdaBoost is firstly applied to only select top hundreds of discriminative features and then another classifier (SVM or ELM) is trained on these selected features. As feature dimension is reduced after AdaBoost being applied, the classifier combination method can also reduce the testing time at the same time. The disadvantage of this method is that the training time increases because it has two training processes.

4. Experiments and analysis

4.1. Real-world smile database

Experiments are conducted on GENKI-4K database, which is the only public database for the study of smile detection in unconstrained scenarios. Some example images in it are illustrated in Fig. 6 and its important properties are listed as below:

(1) This database contains 4000 face images (1838 "non-smile" and 2162 "smile"); (2) The pose range (yaw, pitch, and roll of parameters of the head) of most images is within approximately $\pm 20^\circ$ of frontal; (3) GENKI-4K also has various imaging conditions including, e.g., gender, age, ethnicity, glasses, facial hair, partial occlusion (very few).

4.2. Experimental settings

There are five different baseline features having been compared with GSS feature. Parameter settings of them are listed as below:

Raw pixel: All of the pixel-intensities in the normalized gray-scale image could be concatenated into a feature vector, which is exactly the raw pixel feature.

Gabor: The parameters of Gabor feature are composed of 8 orientations and 5 spatial frequencies (9:36 pixels per cycle at 1/2 octave steps). We downsample the 40 Gabor Energy Filters by a factor of 4, so the Gabor feature vector has 23,040 dimensions.

LBP: For extracting LBP feature, each face image of size of 48×48 is firstly divided into 4×4 sub-regions, each of which is in the resolution of 12×12. Then we adopt 59-label LBP (8, 2, $u2$) operator to compute LBP features for each sub-region. Consequently, the LBP vector has 944(16×59) dimensions.

LPQ: For extracting LPQ feature, each face image is also equally divided into sub-regions. All the sub-regions have the same size of 16×16. So LPQ feature of an input image is a 2304-dimensional feature vector.

HOG: Both of HOG36 and HOG31 have the same number of cells (6×6). As a result, for the same input image, HOG36 is a $6 \times 6 \times 36$ feature map and HOG31 is a $6 \times 6 \times 31$ feature map.

GSS: Since GSS feature is extracted on HOG feature, $m = 6$. Every block has 9 cells and $n = 3$. After using formulation (7) and (8), GSS feature is a 576-dimensional vector.

Apart from parameter settings of baseline features, four-fold cross validation is adopted. To put it simply, all face images in GENKI-4K database are divided into four heaps with approximate same ratio between non-smile and smile faces. At each time, select one distinct heap for testing and let the other three heaps be training samples. This procedure is subsequently repeated three times.

Fig. 6. Smile and non-smile examples in GENKI-4K.

Table 1
Smile detection accuracies of two different face alignment strategies using linear SVM with different features.

Accuracy (%)	Face registration approach	
	Eyes-based	Eyes-and-mouth-based
Raw pixel	85.69 ± 0.22	89.21 ± 0.12
Gabor	92.81 ± 0.55	94.03 ± 0.17
LBP	88.91 ± 1.30	91.08 ± 0.40
LPQ	88.98 ± 0.65	91.13 ± 0.66
HOG36	90.78 ± 0.57	92.58 ± 0.47
HOG31	92.66 ± 0.39	94.06 ± 0.82
GSS	89.49 ± 0.26	90.51 ± 0.35

Regarding computing HOG feature, VLFeat[1] is very helpful for us. As for the implementations of classification algorithms, we refer to LIBLINEAR,[2] GML AdaBoost Matlab Toolbox,[3] and Matlab codes of ELM Algorithm.[4]

4.3. Results and analysis

Table 1 shows the performances of different face registration methods under the same classification condition (SVM). By visualizing these data in Fig. 7, it is easy to get that eyes-and-mouth-based face alignment method outperforms the eyes-based face alignment way, which proves that aligning mouth areas is really helpful for constructing a high-performance smile detector. Besides, the performance improvement is more apparent on raw pixel feature than other features.

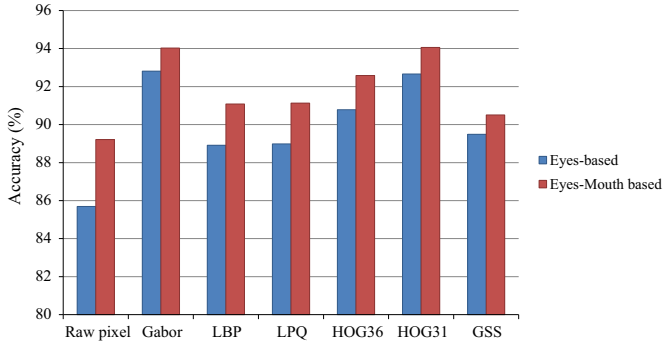


Fig. 7. Smile detection accuracy comparison for different face alignment methods.

In Table 2, the performances of seven different vector comparison functions for GSS feature have been listed. Obviously, the Square Euclidean method achieves the best performance. Therefore, in latter experiments, we extract GSS feature using this method.

As [21] does, we also try the original ELM [33] on GSS feature and other baseline features, the performances of which are illustrated in Fig. 8. A decent number of neurons for all of the features

[1] http://www.vlfeat.org/.
[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/.
[3] http://www.graphics.cs.msu.ru/ru/science/research/machinelearning/adaboosttoolbox/.
[4] http://www.ntu.edu.sg/home/egbhuang/elm_kernel.html.

is around 500. In this figure, GSS feature beats three baseline features, which are LBP, LPQ, and raw pixel. And HOG31 achieves the best performance, but it is still lower than the accuracy of Linear ELM (94.06%).

In order to better understand the distributions of discriminative features when using AdaBoost, we also visualize the selected HOG and GSS features in Figs. 9 and 10 separately. In Fig. 9, discriminative HOG31 features mainly aggregate in the mouth region. In Fig. 10, 13 of 16 images have yellow blocks on the mouth area, which again reinforces the conclusion that mouth region is more important than other places for smile detection problems.

Table 2
The performances of GSS feature using different vector comparison methods under the same classification condition (linear SVM).

Method	Accuracy
Cityblock	89.96 ± 0.53
Chebychev	87.26 ± 0.93
Cosine	89.24 ± 0.93
Correlation	89.68 ± 0.44
Spearman	90.01 ± 0.37
Euclidean	89.81 ± 0.53
Square Euclidean	$\mathbf{90.51 \pm 0.35}$

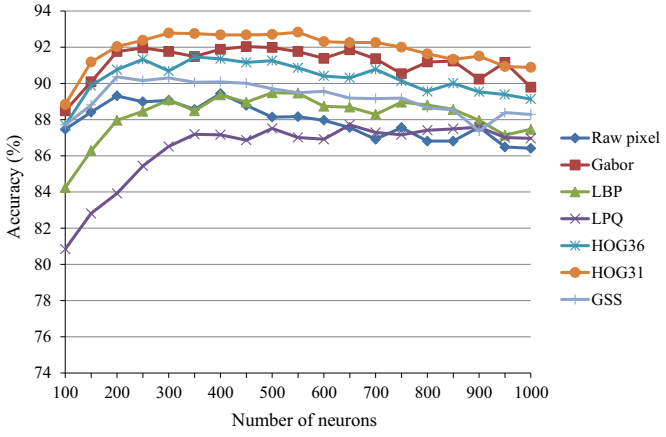


Fig. 8. Smile detection accuracy when using ELM with different numbers of neurons.

Table 3 demonstrates the smile recognition rates of our proposed GSS features with baseline methods. In the same feature condition, classifier combination is better than using any classifier method alone and AdaBoost performs the worst. In the same classifier condition, HOG31 ≈ Gabor > HOG36 > LPQ ≈ LBP > GSS > Raw pixel. We could conclude that (1) Although HOG31 and Gabor perform comparably, Gabor is much more time and space consuming than HOG31; (2) GSS feature beats one of the baseline features, which implies its effectiveness for real-world smile detection tasks. In our previous work [34], HOG31 with GSS already achieves improved performance. Here, we add raw pixel feature into them because it is simple and meanwhile has origin image information. Finally, using HOG31+GSS+Raw pixel with AdaBoost+Linear ELM achieves the best smile detection rate (94.61%). Note that for input images with a same resolution, the time complexity of Raw pixel is $O(1)$, the time complexity of HOG31 is $O(m)$, and the time complexity of GSS is $O(m \cdot n^2)$. Here, m stands for how many cells in a cell-based HOG31 feature map and n stands for how many cells in a block-based HOG31 feature map. Therefore, when using the combined features (HOG31+GSS+Raw pixel), it is better to choose a minor n.

5. Conclusions

This paper mainly focuses on smile detection in unconstrained scenarios. The primary contribution for solving this problem is that a new type of feature, GSS, is proposed. GSS is inspired by CSS in pedestrian detection and shows its effectiveness in smile detection in unconstrained scenarios by experiments. Additionally, aligning the mouth position into fixed place for all the face images is found to be useful for constructing a high-performance practical smile detector. Furthermore, feature combination and classifier combination strategies are utilized in this work. Experiments show that our combined multiple features (HOG31+GSS+Raw pixel) with combined multiple classifiers (AdaBoost+ Linear ELM) achieve the best performance (94.61%) on the GENKI-4K database.

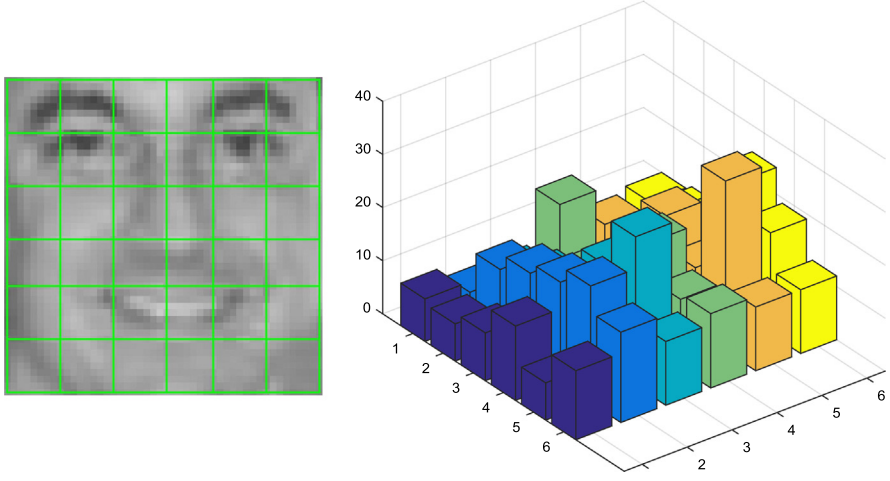


Fig. 9. Top 500 feature positions of HOG31 selected by AdaBoost.

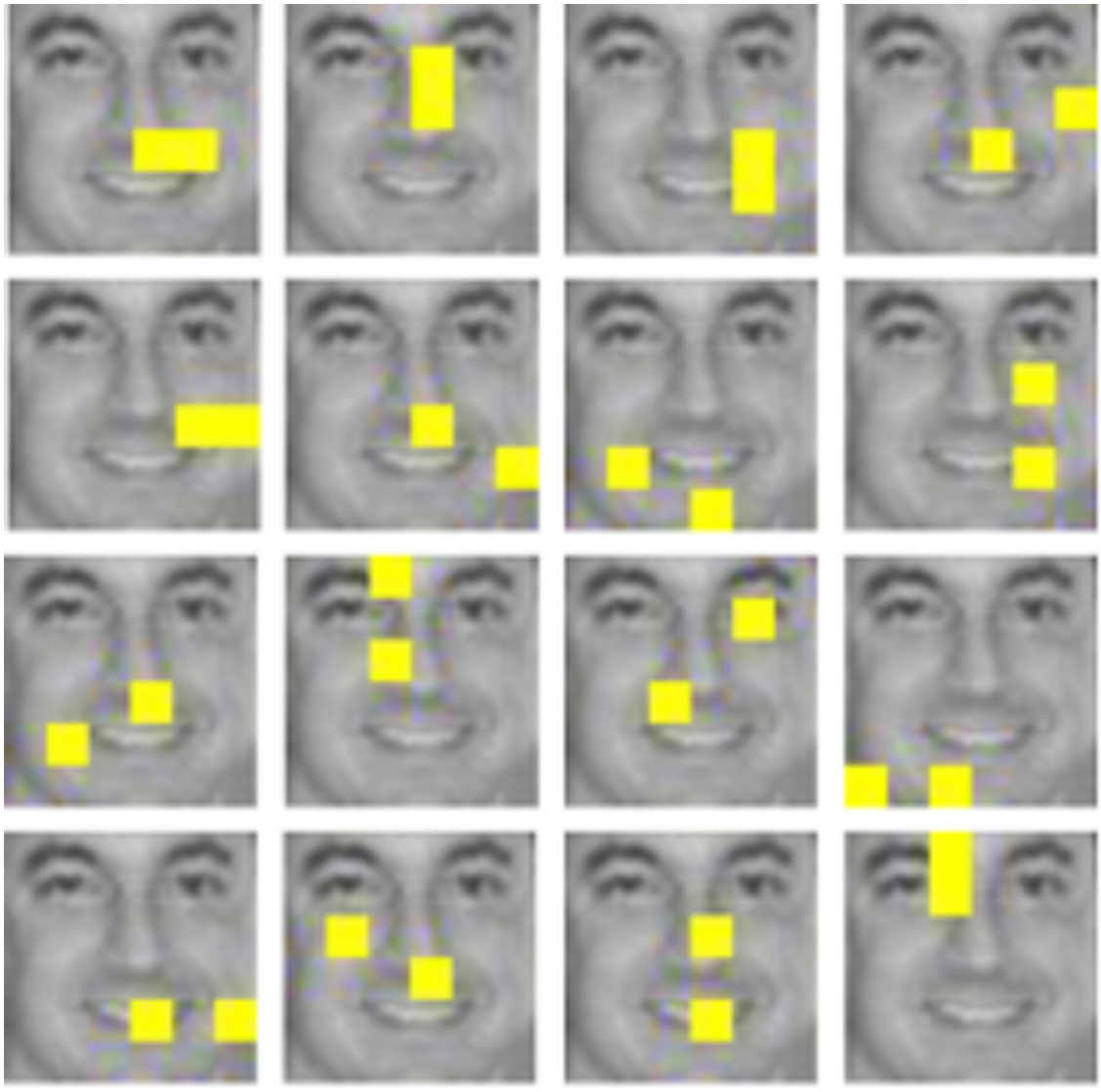


Fig. 10. Top sixteen feature positions of GSS selected by AdaBoost, and they are arrayed row by row and from left to right.

Table 3
Experimental results of smile detection using different features and classifiers.

Accuracy (%)	Classifier				
	AdaBoost	Lin_SVM	Lin_ELM	Ada+Lin_SVM	Ada+Lin_ELM
Raw pixel	88.83 ± 0.43	89.21 ± 0.12	88.44 ± 0.98	**89.49 ± 1.15**	88.83 ± 1.59
Gabor	93.41 ± 0.68	**94.03 ± 0.17**	93.91 ± 0.42	92.88 ± 1.23	92.31 ± 0.47
LBP	89.94 ± 0.43	91.08 ± 0.40	91.26 ± 0.59	91.51 ± 0.11	**91.83 ± 0.82**
LPQ	87.96 ± 0.72	91.13 ± 0.66	91.71 ± 0.60	91.41 ± 0.90	**91.76 ± 1.33**
HOG36	90.88 ± 0.22	92.58 ± 0.47	92.68 ± 0.64	**93.13 ± 0.77**	93.11 ± 0.47
HOG31	92.48 ± 0.61	94.06 ± 0.82	94.06 ± 0.61	94.46 ± 0.78	**94.48 ± 0.65**
GSS	89.01 ± 0.67	90.51 ± 0.35	90.56 ± 0.17	**91.11 ± 0.47**	90.91 ± 0.57
HOG31+GSS+Raw pixel	92.51 ± 0.40	94.28 ± 0.60	94.21 ± 0.35	94.56 ± 0.62	**94.61 ± 0.53**

References

[1] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, J. Movellan, Toward practical smile detection, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 31 (11) (2009) 2106–2111.

[2] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1) (1991) 71–86.

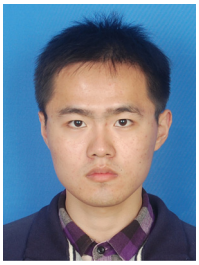
[3] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 19 (7) (1997) 711–720.

[4] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Automatic recognition of facial actions in spontaneous expressions, J. Multimed. 1 (6) (2006) 22–35.

[5] J. Whitehill, C.W. Omlin, Haar features for facs au recognition, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2006, pp. 5–101.

[6] A. Hadid, M. Pietikainen, T. Ahonen, A discriminative feature space for detecting and recognizing faces, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004, pp. 797–804.

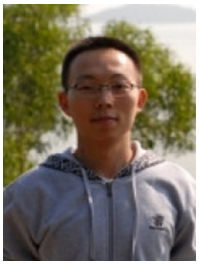
[7] V. Ojansivu, J. Heikkila, Blur insensitive texture classification using local phase quantization, in: International Conference on Image and Signal Processing, 2008, pp. 236–243.
[8] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886–893.
[9] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 32 (9) (2010) 1627–1645.
[10] S. Walk, N. Majer, K. Schindler, B. Schiele, New features and insights for pedestrian detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 1030–1037.
[11] M.F. Valstar, Automatic Facial Expression Analysis, in: Understanding Facial Expressions in Communication, Springer, India, 2015, pp. 143–172.
[12] M.F. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, IEEE Trans. Syst. Man Cybern. Part B: Cybern. (TSMC) 42 (1) (2012) 28–43.
[13] P. Lucey, J.F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, K.M. Prkachin, Automatically detecting pain in video through facial action units, IEEE Trans. Syst. Man Cybern. Part B: Cybern. (TSMC) 41 (3) (2011) 664–674.
[14] S. Yang, B. Bhanu, Understanding discrete facial expressions in video using an emotion avatar image, IEEE Trans. Syst. Man Cybern. Part B: Cybern. (TSMC) 42 (4) (2012) 980–992.
[15] A. Savran, B. Sankur, M.T. Bilge, A. Savran, B. Sankur, M.T. Bilge, Regression-based intensity estimation of facial action units, Image Vis. Comput. (IVC) 30 (10) (2012) 774–784.
[16] Y. Shinohara, N. Otsu, Facial expression recognition using fisher weight maps, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2004, pp. 499–504.
[17] Y. Bai, L. Guo, L. Jin, Q. Huang, A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition, in: IEEE International Conference on Image Processing (ICIP), 2009, pp. 3305–3308.
[18] C. Shan, An efficient approach to smile detection, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2011, pp. 759–764.
[19] C. Shan, Smile detection by boosting pixel differences, IEEE Trans. Image Process. (TIP) 21 (1) (2012) 431–436.
[20] Y. Zhang, L. Zhou, T. Sun, A novel approach to detect smile expression, in: IEEE International Conference on Machine Learning and Applications (ICMLA), 2012, pp. 482–487.
[21] L. An, S. Yang, B. Bhanu, Efficient smile detection by extreme learning machine, Neurocomputing 149 (2015) 354–363.
[22] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 fps via regressing local binary features, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1685–1692.
[23] Z. Zhang, P. Luo, C. L. Chen, X. Tang, Facial landmark detection by deep multi-task learning, in: European Conference on Computer Vision (ECCV), 2014, pp. 94–108.
[24] B. Martinez, M.F. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression-based facial point detection, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 35 (5) (2013) 1149–1163.
[25] M.Uřičář, V. Franc, V. Hlaváč, Detector of facial landmarks learned by the structured output svm, in: International Conference on Computer Vision Theory and Applications (VISAPP), 2012, pp. 547–556.
[26] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1867–1874.
[27] C. Vondrick, A. Khosla, T. Malisiewicz, A. Torralba, HOGgles: visualizing object detection features, in: IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1–8.
[28] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (21) (1997) 119–139.
[29] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Ann. Stat. 28 (2) (2000) 337–407.
[30] R. Lienhart, E. Kuranov, V. Pisarevsky, Empirical analysis of detection cascades of boosted classifiers for rapid object detection, in: Lecture Notes in Computer Science, 2003, pp. 297–304.
[31] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[32] G.B. Huang, An insight into extreme learning machines: random neurons, random features and kernels, Cogn. Comput. 6 (3) (2014) 376–390.
[33] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (2006) 489–501.
[34] H. Liu, Y. Gao, P. Wu, Smile detection in unconstrained scenarios using self-similarity of gradients features, in: IEEE International Conference on Image Processing (ICIP), 2014, pp. 1455–1459.



Yuan Gao was born in Jiangsu, China, in 1990. He received his B.E. degree in Intelligent Science and Technology from Xidian University in 2012. Then he obtained his M.S. degree in Computer Applied Technology from Peking University in 2015. His research interests include object detection, facial expression recognition and facial gender identification.

Hong Liu received a Ph.D. degree in Mechanical Electronics and Automation in 1996, and serves as a Full Professor in the School of EECS, Peking University (PKU), China. Prof. Liu has been selected as Chinese Innovation Leading Talent supported by "National High-level Talents Special Support Plan" since 2013. He is also the Director of Open Lab on Human Robot Interaction, PKU. His research fields include computer vision and robotics, image processing, and pattern recognition. Dr. Liu has published more than 150 papers and gained Chinese National Aero-Space Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors in PKU. He is an IEEE member, vice president of Chinese Association for Artificial Intelligent (CAAI), and vice chair of Intelligent Robotics Society of CAAI. He has served as keynote speakers, co-chairs, session chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC and IIHMSP, recently also serves as reviewers for many international journals such as Pattern Recognition, IEEE Transactions on Signal Processing, and IEEE Transactions on PAMI.

Pingping Wu received a B.E. degree in Information and Computing Science from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009. Currently, she is working toward a Ph.D. degree at the School of Electronics Engineering and Computer Science (EECS), Peking University (PKU). Her current research interests are facial expression recognition, smile analysis, visual speech recognition. Related papers have been published on ICRA, ICIP, ICPR and ICASSP.



Can Wang was born in Shandong, China. He is a doctor student majoring in Computer Applied Technology from Peking University from 2011. His mentor is Prof. Dr. Hong Liu. His research interests include RGB-D motion detection, multiple-tracking and virtual reality.