Neurocomputing 235 (2017) 245–254

Contents lists available at ScienceDirect

Neurocomputing

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

How do you smile? Towards a comprehensive smile analysis system

Pingping Wu[a,b], Hong Liu[a,*], Chao Xu[b], Yuan Gao[c], Zheyuan Li[a], Xuewu Zhang[a]

[a] Key Laboratory of Machine Perception (Ministry of Education), Peking University, Shenzhen Graduate School, China
[b] School of Technology, Nanjing Audit University, China
[c] Department of Computer Science, Christian-Albrechts University, Kiel, Germany

ARTICLE INFO

Communicated by Z. Wang

Keywords:
Smile detection
Intensity estimation
Spontaneous versus posed (SVP)
Facial landmark localization

ABSTRACT

To better understand the expression of human smile, there have been considerable studies about automatic smile detection. Despite all the research, few attention is paid to analyze a smile in a comprehensive way. In this paper, a smile analysis system is presented to detailedly measure a person's smile, which consists of three main modules: smile detection, smile intensity estimation and spontaneous versus posed (SVP) smile recognition. Firstly, our recent proposed feature, Self-Similarity of Gradients (GSS), is employed to detect smiling facial images in unconstrained scenarios. Secondly, the smile intensity is estimated in terms of different facial regions rather than merely the mouth region, which is also applied in the temporal phase segmentation of a smile. Finally, in SVP smile recognition module, a discriminative learning model (DLM) is proposed based on a local spatial-temporal feature, which devotes to obtaining most robust and discriminative patterns of interest. The first two modules are the bases of the last, preparing a deeper understanding of a smile. Experiments on benchmark databases are carried out and compared with the state-of-the-art methods respectively, which validate the advantages of our approach of SVP smile recognition. Moreover, a comprehensive analysis of human smile is given for the first time to the best of our knowledge, which could pave the way for computers that better assess the emotional states of their users and provide useful and important information in helping the research of psychology and behavior science.

1. Introduction

Smile is an influential biometric cue in social interactions, which is also the easiest facial expression to be voluntarily posed [1]. To make a computer better understand human smile, first, substantive smiling facial images or sequences need to be collected, which can be completed by smile detection. Besides, as a fundamental and crucial step of smile analysis, automatic smile detection in unconstrained scenarios provides input data for the following further analyses: smile intensity estimation and spontaneous versus posed (SVP) smile recognition. As is known, smile occurs with different intensities such as “grin” and “chortle”, thus, an effective method of smile intensity estimation can quantify the expression so as to determine the extent of joy. However, smile differs not only in intensity but also in motivation, which brings the diversity of smile. Smile can signal enjoyment, politeness or even can cover other emotions like embarrassment, fear or frustration [2]. So far, 18 different types of simile are identified by Ekman [1] who claimed there might be as many as 50.

A comprehensive smile analysis system is built in this paper, which has wide applications such as interactive systems, video conferences, digital video cameras and patient emotion monitoring. For example, the intensity estimation module is a measuring tool in analysing smile, which can be embedded in video cameras in order to capture the most brilliant smile. People suffering from autism [3] can use SVP smile recognition in social communication to distinguish genuine and deceptive facial expressions. By employing this technique in a video camera, a natural and unforced smile can be captured excluding artificial and posed ones.

In this paper, a comprehensive smile analysis system is proposed, in which three modules are involved including smile detection, smile intensity estimation and SVP smile recognition. Fig. 1 shows a brief framework of our system. First, the basic module of our system is smile detection, through which substantive data of smiling faces can be acquired. To meet the complexity of real-world conditions, smile detection in unconstrained scenarios is focused. The study of this module is explicitly presented in our recent work [4] which won't be repeated here. Second, smile intensities are estimated in terms of different facial subregions rather than merely the mouth region, in which the state-of-the-art method of facial landmark localization is employed. Also, with the computed intensities of different facial

* Corresponding author.
E-mail addresses: pingpingwu@pku.edu.cn (P. Wu), hongliu@pku.edu.cn (H. Liu), yuan.gao@stu.uni-kiel.de (Y. Gao).

http://dx.doi.org/10.1016/j.neucom.2017.01.020
Received 3 March 2016; Received in revised form 21 December 2016; Accepted 10 January 2017
Available online 13 January 2017
0925-2312/

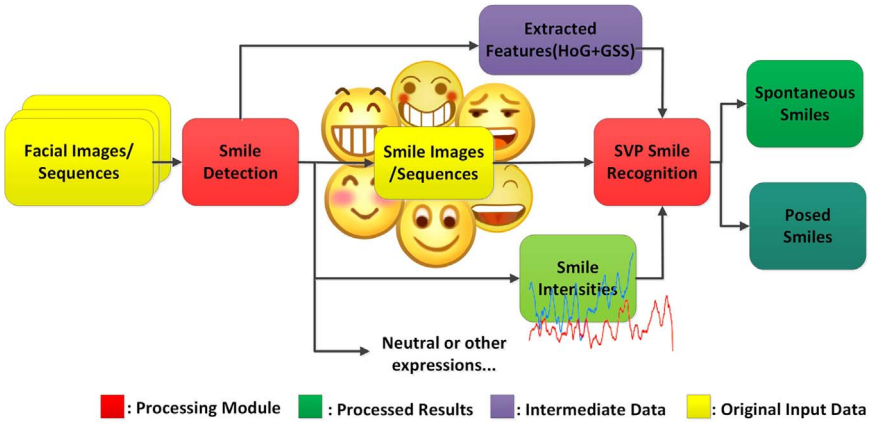


Fig. 1. Brief framework of our smile analysis system.

subregions, temporal phase segmentation of smile is accordingly carried out. Third, most proposed methods for SVP smile recognition extract only geometry-based features [5–7], which may lead to the loss of texture information. In fact, texture information plays an important role for smile analysis, especially for SVP smile recognition. Therefore, an appearance feature based on a discriminative learning model (DLM) is proposed by improving the discriminability of Completed Local Binary Patterns from Three Orthogonal Planes (CLBP-TOP). Note that the third module SVP smile recognition is a refined and expanded version of our conference proceedings paper [8]. In this paper, more details are given about this module and different methods of spatio-temporal division are applied.

The rest of paper is organized as follows: Related work is introduced in Section 2. Section 3 presents smile intensity estimation including facial landmark localization and temporal phase segmentation. SVP smile recognition is addressed in Section 4 where the discriminative learning model is also described. Section 5 demonstrates experimental results and analysis, then conclusions of the paper are drawn in Section 6.

2. Related work

Speaking of automatic smile detection in unconstrained scenarios, the work of Whitehill et al. is the foundation, which is motivated by applying automatic smile detection on commercial digital cameras [9]. Bai et al. proposed a pyramidal representation of HOG (PHOG) for smile detection and achieved comparable performance with regard to Gabor [10]. Shan presented a novel smile detection approach by simply comparing the intensities of a few pixels in an image and achieved better performance than Gabor+SVM [11]. For smile detection, efficient image registration and feature representation are both important [9]. For image registration, works in [12–14] involve common approaches. However, how many facial landmarks contribute to the best smile detector has not been much investigated. For feature representation, there are some traditional feature extraction methods such as PCA [15], LDA [16], Gabor [17,18], Haar [19], LBP [20] and HOG [21]. Recently, a low-dimensional HOG feature is obtained by Felzenszwalb et al., using an analytic dimensionality reduction approach [22]. In our recent work [4], we employed the improved HOG due to its low dimensions and outstanding performance. Besides, based on the observation of HOG's visualization and the inspiration from self-similarity on color channels (CSS) [23], a new descriptor named Self-Similarity of Gradients (GSS) is proposed to capture pairwise similarities of localized gradient distributions. Moreover, an eye-mouth based alignment is adopted in the face registration procedure, which is proven to be more effective than the eye-only based alignment.

Facial expression intensity estimation is a nontrivial task. The majority of automatic approaches characterising intensity use a two-level model known as onset-apex-offset. In Facial Action Coding System (FACS), five intensity levels for each facial action unit (AU) [24] are specified. As mentioned in [25], there is much less work in the literature on AU intensity estimation. Savran et al. [25] comparatively investigated person-independent intensity estimation of 25 AUs, which showed that their proposed intensity estimator based on regression of appearance features proves to be superior to that based on SVM margins. Delannoy et al. [26] proposed an automatic estimation of the dynamics of facial expression using a three-level model (High, Medium and Low) of intensity. They verified that using the FACS intensity scoring led to a considerable overlap between the estimated intensities while using a three level model enabled to classify the intensities with significantly greater degree of accuracy. As presented in work [6,27], the smile intensity is represented using the amplitude of lip corners, where the facial fiducial points must be localized and tracked first.

The task of SVP smile recognition is challenging as posed smiles often look very similar to spontaneous ones, which makes them difficult to distinguish using human eyes. However, they are actually different because they are brought about by different brain systems. From the knowledge of neurology, the face is innervated by two different brain systems that compete for control of its muscles [28]. One is the cortical brain system related to voluntary and controllable behaviors. The other is the sub-cortical systems taking in charge of involuntary expressions. Facial expressions mediate by these two systems show differences both in morphology and dynamics. Accordingly, posed smiles are innervated by the cortical brain system while spontaneous smiles are stimulated by the sub-cortical system. From researches in [29,30], facial expressions initiated by sub-cortical system tend to be more symmetrical, consistent and reflex-like. However, facial expressions initiated by cortical system tend to be less smooth and have more variable dynamics. Accordingly, all types of smiles can be divided into two categories: voluntary (deliberated/fake/posed) smiles and involuntary (spontaneous /genuine/felt) ones.

To date, spontaneous versus posed (SVP) smile recognition has gained certain attention [31–35]. Five markers are proposed by Frank et al. [36] to differentiate the genuine smiles from the posed ones including two morphologic ones: D-maker and symmetry and three dynamic ones: smoothness, duration and synchrony. According to the D-maker, one of the differences between genuine and posed smiles is that when a person really feels happy, zygomatic major contracts together with orbicularis oculi, which is frequently used by common people to distinguish the two types of smiles. However, later studies show that the D-marker is not accurate as the zygomatic major and the orbicularis oculi can both be activated under spontaneous and posed conditions [37] and there is no significant difference in symmetry between posed and spontaneous smiles [32]. Similar conclusions can

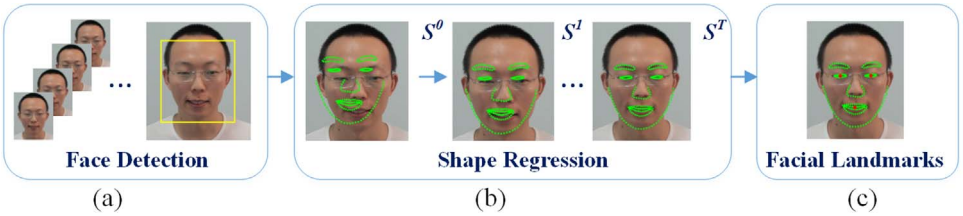


Fig. 2. Process of facial landmark localization (a): input facial sequences then perform the coarse face detection; (b): fast shape regression in a coarse to fine manner; (c): face alignment (The person in the image is a member from our lab).

also be found in [6]. Cohn et al. observed that posed smiles are of larger amplitude and have a less consistent relationship between amplitude and duration than spontaneous ones [31]. Besides, a deceit detection in facial expressions was implemented by Zhang et al. [38], where the enjoyment expression is involved. In order to carry out the detection, they used distance based features and texture based features, achieving an accuracy of 73.16% in deceit detection of enjoyment expression. M. Valstar et al. proposed a method to distinguish SVP smiles by fusing head, face, and shoulder modalities [7]. E. Hoque et al. explored temporal patterns to distinguish delighted smiles from frustrated smiles with their self-built database, achieving the best accuracy of 92% [5]. Moreover, they found that acted instances were much easier for computer to classify than natural ones. H. Dibeklioğlu et al. proposed a geometry-based feature to spot spontaneous and posed smiles and carried out assessment on different facial regions, feature selection, fusion strategies, etc. [6]. Besides, a corpus named UvA-NEMO is collected including 1240 video sequences which is the largest SVP smile database publicly retrievable. Recently, they extended their work by adding gender and age effect analysis and expanded experiments with different classifiers on several databases, more details of which can be obtained in [27]. Pfister et al. proposed a spatiotemporal method to distinguish between SVP facial expressions using the image sequence as a volume with a corpus including both natural and infrared face videos [39]. Instead of extracting geometry feature, they extended the local appearance descriptor into the spatial-temporal descriptor to implement the task. In our earlier work [40], a smile deceit detection has been done by training AU6 and AU12 simultaneously on a static-image database. It should be noted that subtle smiles appeared in a fleeting time as micro-expressions are not considered in the research scope of spontaneous versus posed (SVP) smile recognition. For more details about micro-expressions, refer to the specific research topic as micro-expression recognition [41].

We summarize our main contributions of the paper as follows. 1) It is the first time that a comprehensive analysis of human smile is given to the best of our knowledge including smile detection, smile intensity estimation and SVP smile recognition. 2) DLM is proposed to obtain discriminative and robust patterns extracted by CLBP-TOP. 3) Taking into account of the non-synchronous motion of different facial regions, temporal segmentation is implemented according to the corresponding facial region.

3. Smile intensity estimation

As the module of smile detection is presented in detail in our recent work [4], we skip the module and introduce the module of smile intensity estimation. In this paper, we follow the work in [6,27], where the smile intensity is computed using the amplitude of lip corners. Therefore, facial landmarks like lip corners, eye or eyebrow corners are crucial for the intensity estimation. In order to obtain accurate facial landmarks, an efficient method of facial landmark localization needs to be explored. In addition, geometry-based feature extraction commonly relies on accurate and reliable detection and tracking of fiducial points. If facial landmarks can be precisely located and tracked, the accuracy of geometric feature extraction will be promoted. Therefore, attentions are paid to the state-of-the-art methods of facial landmark localization. Recently, great improvements have been made in the research field [42–45], where facial landmarks can be located faster and more accurately than traditional active appearance model (AAM) [46]. Inspired by these works, a method of shape or semantic facial landmarks is adopted without using any parametric shape models proposed in [43], which has shown extraordinary performance in both accuracy and efficiency. In the following, we firstly introduce the shape regression model as shown in Fig. 2.

3.1. Facial landmark localization

Different from [43], a smart restart approach [44] is added to predict failure cases early on. In addition, a two-level cascaded regression and a correlation-based feature selection are adopted similar to [43,44]. Concretely, for N facial landmarks $S = [x_1, y_1, \ldots, x_N, y_N]^T$ as a face shape, the goal is to find a shape S that is as close as possible to the true shape $\hat{S}$, i.e., minimizing $\| S - \hat{S} \|$. Specifically, ferns are chosen as the primitive regressors. A fern is a classification based regressor that takes an F dimensional input feature vector and computes an output vector by classifying the input into one of the 2^F bins. The classification is performed by comparing the input feature vector to F thresholds of the fern. The classification result is an F dimensional binary vector $f = (f_1, f_2, \ldots, f_F)$ where f_i=0 if the i-th attribute is greater than the corresponding threshold and f_i=1 otherwise. The output of the regressor is determined by looking up the output entry for the classification vector.

Our algorithm generates face shape by repeatedly refining an initial guess shape via a series of cascaded regression functions. The refinement using a regression function R is regarded as a stage, and there are in total T stages. In each stage t, the output of the previous stage S^{t-1}, together with the input image I are used to predict a difference shape ΔS. The sum of the difference shape and the current predict shape is the output of stage t, which can be formulated as:

$$S^t = S^{t-1} + R^t(I, S^{t-1}), \quad t = 1, \ldots, T. \tag{1}$$

wheare $t \in [1..T]$ is the stage index. Given N training examples $\{(I_i, \hat{S}_i)\}_{i=1}^N$, the regressors are sequentially learnt until the training error no longer decreases:

$$R^t = \arg\min_R \sum_{i=1}^{N} \| \hat{S}_i - (S_i^{t-1} + R(I_i, S_i^{t-1})) \| \tag{2}$$

where I_i is a facial image and S_i^{t-1} is the estimated shape in previous stage.

The regression functions are represented by a series of weak regressors, i.e.

$$R^t = (r_1, r_2, \ldots, r_k) \tag{3}$$

where each r_j is called a primitive regressor. In each stage, the input shape is refined by all the primitive regressors in a cascaded manner. As shown in Fig. 3, k primitive regressors form a chain of regressors, where each primitive regressor is called a level. The level k regressor

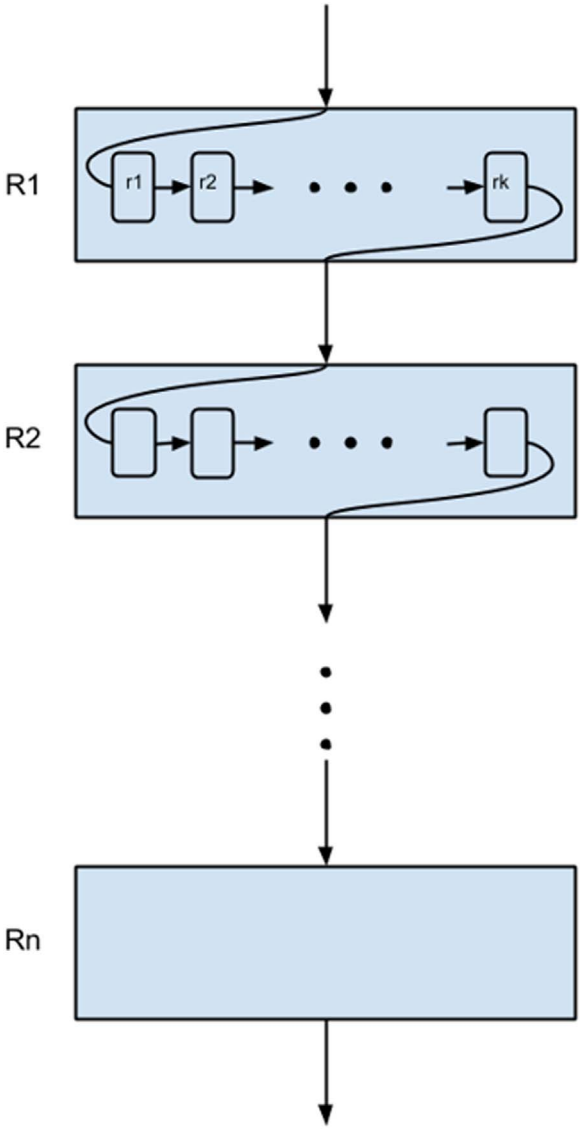


Fig. 3. Illustration of the regression function. A regressor function has n stages, and each stage contains k primitive regressors. The guess shape is refined by all primitive regressors in a cascaded manner.

takes the output shape of previous level as input and predict a shape vector to refine the input shape vector. The refined shape vector is then passed to the next level of regressor for further refinement.

Typically hundreds of primitive regressors are used in each stage. The primitive regressors are weak regressors because they are only able to reduce the shape error slightly, therefore a series of primitive regressors are needed in each stage. Although each primitive regressor alone is not able to reduce the shape error much, all the primitive regressors collectively are able to reduce the shape error significantly. With enough number of primitive regressors, the final regression function becomes very powerful.

Since the output of the algorithm is a refinement of the input shape, the quality of the initial guess would inevitably affect the alignment output. To remove the effect of this randomness, multiple initial shapes are used simultaneously and the final shape is computed as a combination of all outputs.

In the preprocessing of regression, a rough face box is detected, then the landmark is estimated in a coarse-to-fine way. Next geometric centers of eyes and month can be detected, which can be employed for the facial image alignment. Using the presented facial landmark

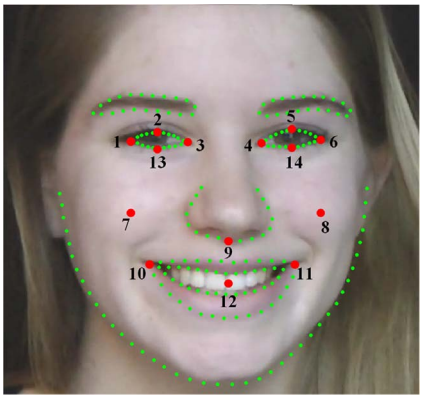


Fig. 4. Applied landmarks for intensity estimation and geometric feature extraction.

localization method, up to 194 green facial landmarks can be located as shown in green in Fig. 4. For intensity estimation, 11 landmarks are reserved while three extra points are added which are the mouth center point 12, lower eyelid center point 13 and 14. All the employed landmarks are shown in red as depicted in Fig. 4.

3.2. New definition of smile intensity

In [2], it has been shown that the timing of spontaneous and posed smiles is different. Instead of dividing time axis into equal length [47], a temporal segmentation method according to smile phases (rise, sustain and decay) is employed, which is defined by M. Hoque et al. to better analyze smile dynamics [48]. And, this segmentation method is proved to be more reasonable for smiles often have a sustained region with multiple peaks, thus there is often not one clear apex or peak to the smile.

In work [27], smile temporal phases are segmented only according to the intensity of lip corner movements. However, movements of lip corners (point 10 and 11), cheek centers (point 7 and 8) and eyelid centers (point 2, 5, 13, and 14) may not rise or decay simultaneously. With the problem considered, temporal phase segmentation is implemented independently for each facial subregion. Since accurate and abundant facial landmarks have been localized in each frame, the intensity of lip corner movement is defined as follows:

$$\mathcal{I}_{lip}(t) = \frac{d(p_{12}^t, p_{10}^t) + d(p_{12}^t, p_{11}^t)}{d(p_{12}^1, p_{10}^1) + d(p_{12}^1, p_{11}^1)} \tag{4}$$

where p_i^t is the location of the i-th landmark in the t-th frame and $d(\cdot)$ denotes the Euclidean distance. Therefore, the longest continuous increase of $\mathcal{I}_{lip}$ is defined as the rise phase for mouth region while the longest decrease of $\mathcal{I}_{lip}$ the decay phase. The sustain phase for mouth region is between the last frame of the rise phase and the first frame of decay phase. Similarly, the amplitude of eyelid movement and cheek center movement are respectively computed by:

$$\mathcal{I}_{eyelid}(t) = \frac{d(p_2^t, p_{13}^t) + d(p_5^t, p_{14}^t)}{d(p_2^1, p_{13}^1) + d(p_5^1, p_{14}^1)} \tag{5}$$

$$\mathcal{I}_{cheek}(t) = \frac{d(p_9^t, p_7^t) + d(p_9^t, p_8^t)}{d(p_9^1, p_7^1) + d(p_9^1, p_8^1)} \tag{6}$$

The temporal segmentation for cheek region is similar to mouth region, where the longest continuous increase of $\mathcal{I}_{cheek}$ is defined as the rise phase for cheek region and the longest decrease of $\mathcal{I}_{cheek}$ the decay phase. The sustain phase for cheek region is between the last frame of the rise phase and the first frame of decay phase. However, the eye region is different from both mouth and cheek region as the eye aperture becomes small with the increase of smile intensity, especially

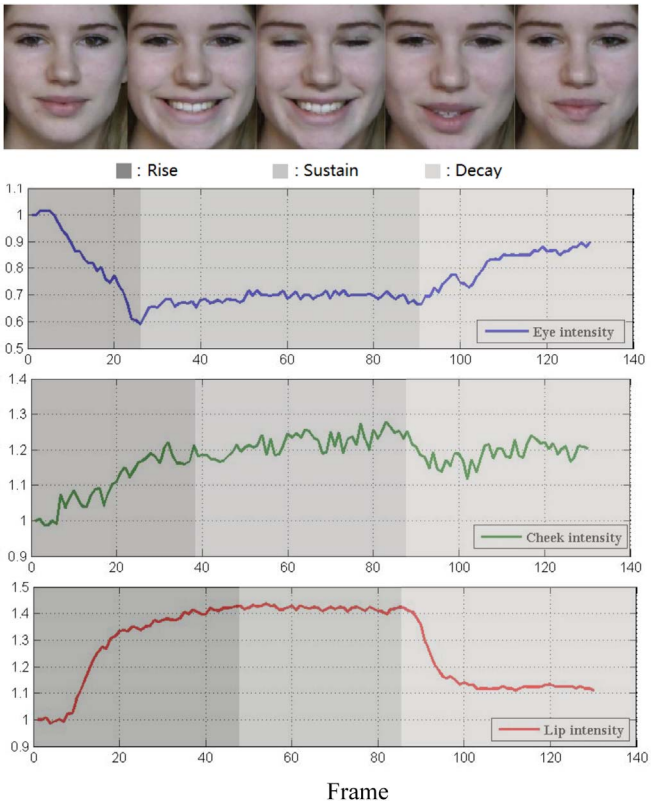


Fig. 5. Intensities and temporal phases for different facial regions.

in the sustain phase. Therefore, the rise phase for eye region is defined as the longest continuous decrease of $\mathcal{I}_{eye}$ while the decay phase the longest increase of $\mathcal{I}_{eye}$. The sustain phase for eye region is between the last frame of the rise phase and the first frame of decay phase. Fig. 5 shows the intensity and temporal segmentation for each facial subregion, which verifies that the same temporal phase for eye, cheek and mouth region may happen in different time.

3.3. Geometric features

With the derived landmarks and our newly defined amplitude/intensity of each facial subregion, geometric features can be extracted through the calculation proposed in [27]. Specifically, five types of geometric features are extracted which are amplitude related, amplitude/duration related, duration related, speed related and acceleration related. For example, the amplitude related involves amplitude ratio, max amplitude, STD amplitude, total amplitude and net amplitude. Speed is the first derivative of amplitude about time while acceleration is the second derivative of amplitude about time. Duration is computed by dividing occupied frames with frame rate. All detailed calculation formulas can be referred in [6].

4. SVP smile recognition

For SVP smile recognition, to catch hold of robust and discriminative features is a key step. In [27], a set of geometric feature is proposed, which shows effective and favorable results. Geometric features represent facial landmark displacements, curvature changes of lips and eyelids, size of eyes, etc. Alternatively, appearance features characterize texture information brought by facial muscle movements like eye corner wrinkles, which is an indispensable and non-substitutable element. Therefore, we focus on exploring a set of texture or appearance-based feature, which can be combined with the sophisticated geometric features to make the recognition more accurate.

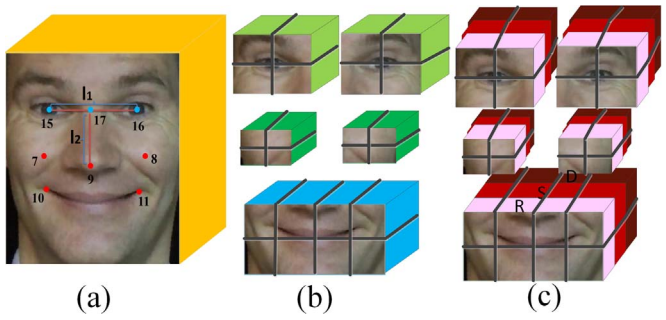


Fig. 6. (a) Facial key points of a subject from UvA-NEMO (b) Cropped subregion volumes (c) Divided blocks in spatial-temporal domain.

4.1. Division in spatiotemporal domain

As our proposed appearance feature by improving completed LBP from three orthogonal planes (CLBP-TOP) is a local descriptor, global information is absent. To overcome the limitation, the whole image sequence is usually equally divided into blocks in both spatial and temporal domain. As previous studies have shown that the subregions such as eyes play important roles for SVP smile recognition [2,49], flexible facial subregion cropping (FSC) is applied considering specific facial regions. Besides, it also takes into account that different subjects' facial organs are of different sizes and changeable when speech and expressions occur. Since facial landmarks have been localized, five landmarks (center of eyes p_{15}, p_{16}, lip corners p_{10}, p_{11}, nose tip p_9) are employed to locate facial subregions as shown in Fig. 6(a). With parameter α_1, α_2, β_1, and β_2 controlling region size according to the prior knowledge of face proportion, facial subregion volumes illustrated in Fig. 6(b) can be derived. As the smile intensity for different facial subregions is defined in the previous section, division in temporal domain for each cropped subregion is accordingly carried out, i.e., each subvolume V can be divided into three blocks as shown in Fig. 6(c).

There are several advantages of FSC: 1) subregions of different sizes tend to gather relevant facial textures and avoid fragmentation of associated information; 2) subregions are flexible since the cropping is implemented according to different subjects' organ sizes; 3) some redundant information could be filtered out such as nose and forehead which are relatively static.

4.2. Discriminative learning model (DLM)

Motivated by finding the optimal subset of patterns and inspired by the idea from [50], a discriminative learning model (DLM) containing three layers is proposed here. Instead of using the original local binary pattern (LBP), Completed local binary pattern (CLBP) shown good performance in texture classification is adopted here [51]. Based on LBP, CLBP computes two other items besides the local difference of sign (S) which are the local difference of magnitude (M) and central pixel intensity (C), which can be represented as follows:

$$CLBPH = [CLBPS, CLBPM, CLBPC] \tag{7}$$

CLBPS and the original LBP are the same, *CLBPM* and *CLBPC* compute the local difference of magnitude and central pixel intensity, respectively. To capture dynamics of the smile process, the purely spatial descriptor needs to be extended to spatial-temporal domain. In [39], CLBP-TOP is proposed by extracting CLBP features from three orthogonal planes. Consider that the feature vector would be very long if concatenating all the histograms computed by CLBP-TOP on each divided block. Therefore, the proposed DLM is applied to adaptively learn effective patterns from the database as illustrated in Fig. 7, which presents in detail as follows:

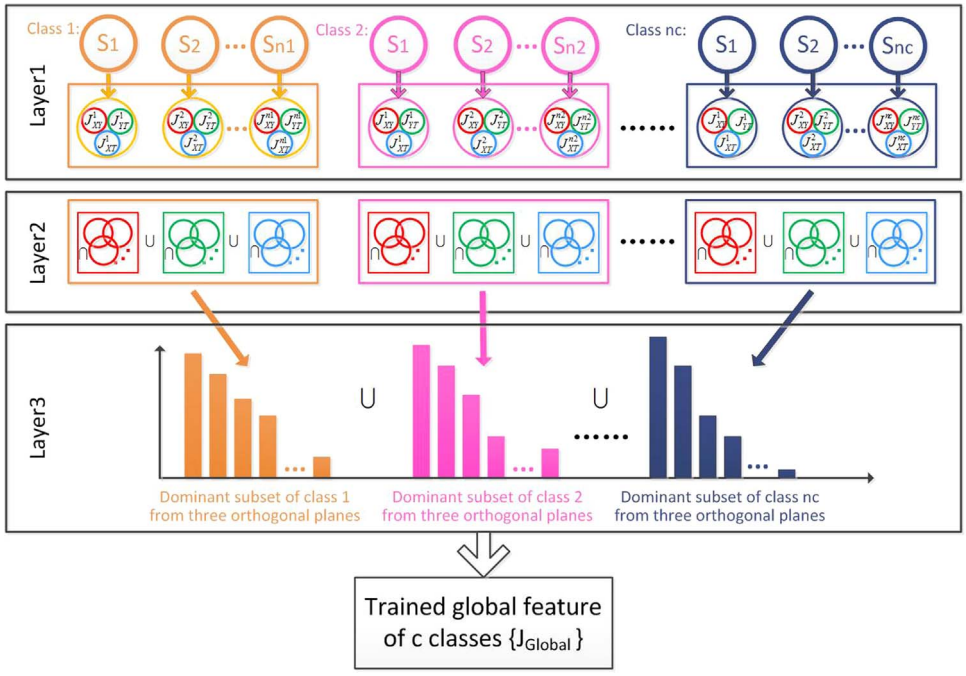


Fig. 7. Discriminative learning model.

Algorithm 1. Learning process of DLM.

Input: class c with n_c examples $\{S_1, S_2, ..., S_{n_c}\}$, B is the total number of blocks, h_i is the histogram of the original pattern sets of each training example S_i, δ is the threshold parameter, $J^i_{u,v}$ is the dominant pattern with respect to each training example S_i in the u-th plane and v-th block;

Output: trained global feature J_{Global} of class c

```
for i = 1 to n_c do
    divide S_i into B blocks, v = 1, ..., B;
    initialize vector P = {P[1], ..., P[p]}, p denotes the total number of pattern types;
    obtain ĥ_i by sorting h_i in descending order;
    obtain P̂ by rearranging vector P according to ĥ_i;
    for k = 1 to p do
        if Σ_{n=1}^{k} ĥ_{i,n} / Σ_{n=1}^{p} h_{i,n} ≥ δ then
            Break
    compute dominant pattern J^i_{u,v} = {P̂[1], ..., P̂[k]}
    J_{u,v} = J_{u,v} ⋂ J^i_{u,v};
for u = 1 to 3 and v = 1 to B do
    J_Global = ⋃_{v=1}^{B} ⋃_{u=1}^{3} J_{u,v} ;
return J_Global ;
```

Layer 1, **Feature Robustness**: Considering feature robustness of texture images, the pattern occurrence needs to be paid close attention. Obviously, frequently occurred patterns of one texture image tend to be more reliable than rarely occurred ones as the rarely occurred patterns can be easily interfered by image noise resulting in a sparse histogram representation. With the above thinking, dominant pattern set is defined here as the minimum set of pattern types covering δ ($0 < \delta < 1$) of all patterns. Denote p_u as the total number of pattern types in the u-th orthogonal plane (u=1: XY, 2: XT, 3: YT plane) and $P_{u,\xi}$ the number of occurrences of pattern type ξ in the u-th orthogonal plane. The dominant pattern set of the u-th orthogonal plane J_u is then calculated as:

$$J_u = arg\min |J_u| s.t. \frac{\sum_{\xi \in J_u} P_{u,\xi}}{\sum_{k=1}^{p_u} P_{u,k}} \geq \delta \tag{8}$$

where $|J_u|$ denotes the number of elements in J_u. Then, the most frequently occurred patterns in each plane can be preserved for the operation of the next layer.

Layer 2, **Feature Discriminability**: In this layer, intra-class variance is taken into account to improve the discriminative power of extracted features. Normally, it is desired that examples belonging to the same class have same dominant pattern set (minimize intra-class variance) while the dominant pattern set of examples from different classes have large difference (maximize inter-class variance). Therefore, intersection of dominant pattern sets is carried out across all training examples in the same class. Thus, the robust and discriminative pattern subset J_c learned from class c with n_c examples can be expressed as:

$$J_c = \bigcap_{n=1}^{n_c} J_u^n \tag{9}$$

where $J^n{}_u$ denotes the dominant pattern set from u-th plane of n-th example. Specifically, according to Eq. (9), applying CLBP-TOP descriptor, the robust and discriminative pattern subset JS_c and JM_c with respect to sign and magnitude component learned from class c with n_c examples can be separately expressed as:

$$JS_c = \bigcap_{n=1}^{n_c} JS_u^n \; JM_c = \bigcap_{n=1}^{n_c} JM_u^n \tag{10}$$

where $JS^n{}_u$ and $JM^n{}_u$ denote the dominant pattern set from u-th plane of n-th example with respect to sign and magnitude component, respectively. Commonly, the central pixel intensity component is not considered for it makes less contribution than the other two components [39,51].

Layer 3 **Feature Representation**: In the previous section, we divide each image sequence into B blocks in spatial-temporal domain. To derive the global feature representation, the extracted robust and discriminative subset J_c in each block B and is then concatenated together as $J_{Global} = \bigcup_{v=1}^{B} J_{c,v}$, where v denotes the number of blocks.

Fig. 8. Frames from UvANEMO database.

The learned J_{Global} from all classes is then put together as the reference for feature extraction of testing sets. Algorithm 1 gives the detail procedure of DLM.

5. Experiments and analysis

5.1. Databases

UvA-NEMO database[1] [6] is so far the largest database for genuine and posed smile recognition, including 597 genuine smile videos and 643 posed ones collected at 50 frames per second with a resolution of 1920×1080 pixels under artificial daylight illuminations. It involves 400 subjects (185 females and 215 males) within an age range from 8 to 76.

SPOS corpus[2] [39] consists of both natural color and infrared videos. Only the onset phase of six basic expressions is recorded with participants' faces cropped out already. Since we focus on genuine smile recognition, only the natural color videos of happy expression are employed here. There are 66 genuine and 14 posed smiles captured with a resolution of 640 × 480 pixels at 25 frames per second in an indoor bunker environment involving 7 subjects (3 females and 4 males).

BBC database is from the "Spot the fake smile" test on BBC website,[3] which consists of 10 genuine and 10 posed smile videos collected with a resolution of 314×286 pixels at 25 frames per second from 7 females and 13 males.

MMI database[4] [52] is not especially collected for SVP smile recognition. In order to compare with work [27], 74 posed smiles from 30 subjects are directly employed and 120 spontaneous smiles of 15 subjects are selected from 383 manually annotated segments. The employed subset of the database is denoted by MMI♯. The database is recorded with two formats: 640 × 480 pixels at 29 frames per second and 720 × 576 pixels at 25 frames per second.

Segments of UvA-NEMO, BBC, and MMI♯start/end with neutral or near-neutral spontaneous or posed smile expressions while image sequences of SPOS only include the onset phase. And some exemplar frames from UvA-NEMO database are illustrated in Fig. 8.

In the preprocessing of facial landmark regression, a rough face box is detected, then the landmark is estimated in a coarse-to-fine way. Next geometric centers of eyes and mouth can be detected. Then the whole face region is normalized with respect to the positions of eyes on which FSC is executed. Different original patterns of CLBP with different radii R=1, 3, and neighboring samples N=4, 8 are tried. CLBP with R=3, N=8 performs best and is denoted as CLBP $_{8,3}$ which is employed as the basic operator in the following experiments. For experiments on UvA-NEMO, two-level 10-fold cross validation is applied: a fold is separated as test set each time, the other 9 folds are used as training sets with cross validation, parameters are optimized without using the test set. In this paper, linear SVM is applied for its simplicity, speed and good performance.

5.2. Evaluation of facial landmarks

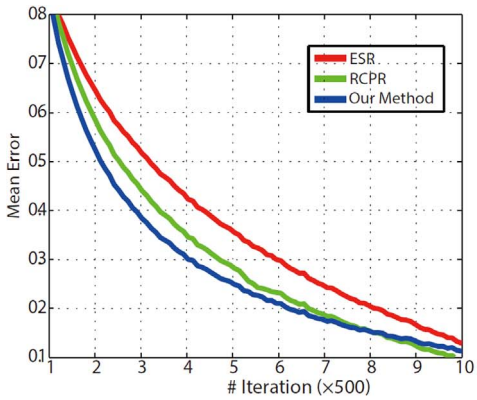
To evaluate the performance of our landmark localization, we report the average error and speed which is measured in frames per second (fps). Errors are measured as the average landmark distance to ground-truth, normalized as percentages with respect to interocular distance [44]. The Helen dataset is employed for the experiment [42]. Here, only the six mouth landmarks are selected to measure the mean error. As shown in Fig. 9, our method has a rapid convergence speed without significantly reducing the accuracy, which is much better for the amplitude calculation.

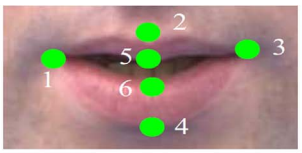
5.3. Comparison of different spatiotemporal divisions

As to space division, FSC is introduced. And the time division based on intensity of different facial subregion is also presented. For FSC, parameters are assigned based on empirical value of comprehensive facial proportion: α_1=α_2=0.8, β_1=0.6, and β_2=0.8. As to time division, our method divides the temporal phase for each facial subregion using the corresponding facial subregion intensity, which is denoted as R-S-D in Table 1. The other two time division methods are considered to compare with R-S-D. One denoted as r-s-d divides the temporal phase for different facial regions barely according to the intensity of the mouth region as introduced in [6,27]. The other equally divides the temporal phase into several parts.

The experiments are implemented on UvA-NEMO using our proposed DLM. In Table 1, $H \times H$ and T EQUAL indicate dividing the whole face into $H \times H$ equal blocks in the spatial domain and T equal blocks in the temporal domain. As to H=1, 2, 4, 8, 10 and T=1, 2, 3, the best accuracy rate 85.26% is achieved with H=8 and T=3, which implies over-dividing and coarse dividing are both undesirable. Over-dividing makes the statistics of local texture invalid while coarse dividing can not well construct the global structure. T is equal to 3 when combined with FSC and H is equal to 8 when combined with R-S-D and r-s-d. FSC performs better than $H \times H$, verifying the mentioned advantages of FSC. Moreover, FSC alleviates influences brought by distributions of redundant information to a certain extent. For different time division methods, R-S-D performs better than r-s-d and T EQUAL when combined with FSC. This verifies that the temporal phase segmentation for eye and cheek are different from the lip. In fact, most of the time, movements in different facial subregions are not simultaneous. R-S-D properly divides the timing of cropped facial subregions according to the corresponding intensity while T EQUAL

[1] [Online] Available http://www.e-nemo.nl.
[2] [Online] Available http://www.ee.oulu.fi/gyzhao/.
[3] [Online] Available http://www.bbc.co.uk/science/humanbody/mind/surveys/smiles/.
[4] [Online] Available http://www.mmifacedb.eu.





Method	Mean Error	FPS
ESR [42]	7.1	2
RCPR [43]	6.5	6
Ours	6.7	13

(b)

Fig. 9. (a) Fast convergence and accurate estimation with no occlusion mechanism and interpolated shape-index feature compares with [44], but has smart restarts compared with [43]. (b) Average error of selected six landmarks in mouth region and speed.

Table 1
Comparison of different space and time division methods using DLM on UvA-NEMO.

Time division	Facial region cropping	Accuracy (%)
R-S-D	FSC	92.03
r-s-d [6,8]	FSC	91.40
T EQUAL	FSC	87.54
R-S-D	$H \times H$	85.87
r-s-d [6,8]	$H \times H$	85.03
T EQUAL	$H \times H$	85.26

ignores this point, directly cutting into several blocks in temporal domain. However, using the time division R-S-D combined with the space division $H \times H$, the improvement is not very obvious. Considering equal division in space domain and then applying time division using R-S-D, the temporal phase of some volumes formed by $H \times H$ can be inaccurately segmented as the time division R-S-D is based on the corresponding intensity of divided subregion using FSC.

5.4. Effect of rise, sustain and decay

To evaluate the effect of different phases of a smile, the DLM feature of each phase are used separately. First, SVM is utilized to classify the feature of each phase individually. Then, the voting strategy is employed to obtain the final classification rate by fusing results of classifiers. As shown in Fig. 10, the decay phase shows lower performance than the other two phase, which is consistent with the fact that when a smile comes to the decay phase, its intensity may appear large fluctuations accompanying with subordinate smiles. Compared to all phases employed R-S-D, the rise phase combined with sustain phase R-S achieves close performance. From the point of extracted DLM features, statistics of the number of robust and discriminative pattern types extracted in the each phase are made. It is found that the number of robust and discriminative pattern types extracted in the decay phase are less than the other two phases. And this may be the essential reason leading to low accuracy in decay phase.

5.5. Comparison with other methods

To evaluate our proposed smile analysis framework, we only need to evaluate the SVP smile recognition module as the first smile detection module is evaluated in [4] while the SVP smile recognition module involves the evaluation of the second smile intensity estimation module. For SVP smile recognition, our method is compared with the state-of-the-art methods on four databases with the same experimental protocols. DLM and CLBP-TOP are implemented with FSC and R-S-D division.

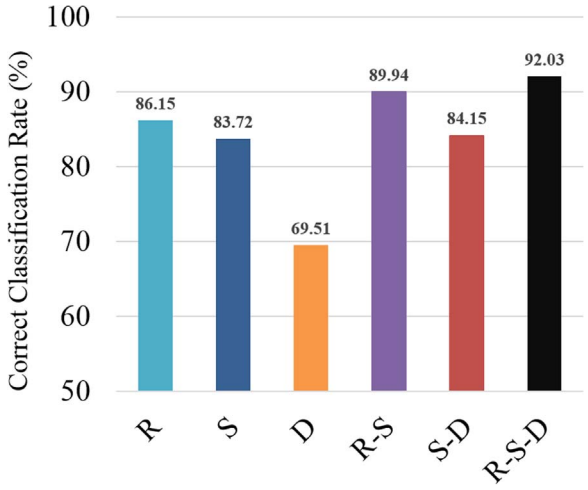


Fig. 10. Evaluation of different smile phases on UvA-NEMO.

In Section 3, the state-of-art facial landmark localization method is introduced, with which the geometric feature (GF) proposed in [6,27] can be extracted. Considering combining the geometric feature with our proposed appearance feature DLM, a feature fusion method needs to be employed. For feature fusion, there are three different fusion strategies of SVM which are early, mid-level and late fusion. Concretely, early fusion means that features from each phase and each classifier are concatenated into one vector and classified by a single SVM classifier; Mid-level fusion denotes that features of all three temporal phases are concatenated for each subregion separately, and the region-based vectors are classified by SVMs; Late fusion individually classifies features from each phase and subregion. For mid-level fusion and late fusion, majority voting strategy is employed which counts the output of each SVM classifier as a single vote and selects the class winning the most votes. The mid-level fusion is applied here as it achieves the best performance.

Correct recognition rates are given in Table 2. DLM achieves better results than CLBP-TOP which validates its discriminative and robust power. The performance of DLM is slightly better than our previous work [8], which is mainly due to the updated time division method based on our new definition of smile intensity. As shown in Table 2, the fusion of DLM and GF outperforms the others on UvA-NEMO, SPOS, and MMI, achieving the highest performance of 94.25%, 81.50%, and 91.06% respectively. It also can be observed that the single DLM shows competitive performance compared to the methods listed below. The recognition rate on SPOS is lower than the other three since SPOS only contains the onset phase of smiles. The recognition accuracy on BBC is

Table 2
Correct recognition rates on four benchmark databases.

Method	Correct Classification Rate (%)			
	UvA-NEMO	BBC	SPOS	MMI♯
DLM+GF (ours)	**94.25**	**90.00**	**81.50**	**91.06**
DLM (ours)	92.03	90.00	79.50	88.31
CLBP-TOP	83.03	80.00	71.5	82.72
Dibeklioğlu et al. (TMM2015) [27]	92.90	90.00	78.75	90.21
Wu et al. (ICASSP2014) [8]	91.40	90.00	79.50	86.10
Dibeklioğlu et al. (ECCV2012) [6]	87.02	90.00	75.00	86.43
Dibeklioğlu et al. [49]	71.05	85.00	66.25	72.55
Pfister et al. (ICCVw2011) [39]	73.06	70.00	67.50	81.37
Cohn et al. [31]	77.26	75.00	72.50	79.02

not improved using our approaches which remains 90.00% as the same as the methods in [6,8,27]. This is due to that BBC and SPOS are of low resolutions, resulting in insufficiency or deficiency of appearance features to some extent. In addition, work [27] is an expansion of work [6], where the recognition accuracy is improved from 87.02% in [6] to 92.90% in work [27] by adding age group information in the feature. Normally, the appearance of a human face naturally changes with the growth of age. Therefore, adding age group features can be regarded as adding appearance information. This is consistent with the experimental results that the best accuracy is attained when combing geometric feature with appearance feature.

6. Conclusions

In this paper, a new comprehensive smile analysis framework is presented that subdivides the smile analysis problem into a cascade of smaller tasks which are smile detection, smile intensity estimation and spontaneous versus posed smile recognition. The work carries forward our recent work of smile detection in unconstrained scenarios. For smile intensity estimation, a new definition is presented according to different facial regions based on the state-of-the-art facial landmark localization method. For SVP smile recognition, we propose a discriminative learning model devoting to obtaining the optimal subset of CLBP-TOP. Experimental results show that the recognition rate is improved using the proposed discriminative learning model, which confirms its robustness and discriminability. Besides, the temporal phase segmentation method based on our new defined smile intensity achieves the best result when comparing with equal temporal division and division only with single mouth intensity, which verifies the effectiveness of the proposed smile intensity estimation. Experiments on four benchmark databases also show our proposed DLM achieves competitive results compared with the state-of-the-art methods. And the best performance 94.25% is achieved when combined with the geometric feature. As the smile expression is recognized in a deep analytic way in our work, it could raise the possibility of inviting many new applications in the future.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC, no. 60675025), the National High Technology Research and Development Programme of China (863 Programme, no. 2006AA04Z247), the Scientific and Technical Innovation Commission of Shenzhen Municipality (nos. JCYJ2013033-1144631730, JCYJ20130331144716089), and the Specialized Research Fund for the Doctoral Programme of Higher Education (SRFDP, no. 20130001110011).

References

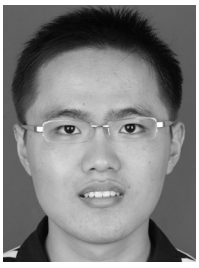
[1] P. Ekman, Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition), WW Norton & Company, 2009.
[2] Z. Ambadar, J.F. Cohn, L.I. Reed, All smiles are not created equal: morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous, J. Nonverbal Behav. 33 (1) (2009) 17–34.
[3] S. Baron-Cohen, H.A. Ring, E.T. Bullmore, S. Wheelwright, C. Ashwin, S. Williams, The amygdala theory of autism, Neurosci. Biobehav. Rev. 24 (3) (2000) 355–364.
[4] Y. Gao, H. Liu, P. Wu, C. Wang, A new descriptor of gradients self-similarity for smile detection in unconstrained scenarios, Neurocomputing 174 (2016) 1077–1086.
[5] M. Hoque, D. McDuff, R. Picard, Exploring temporal patterns in classifying frustrated and delighted smiles, IEEE Trans. Affect. Comput. 3 (3) (2012) 323–334.
[6] H. Dibeklioğlu, A.A. Salah, T. Gevers, Are you really smiling at me? Spontaneous versus posed enjoyment smiles, in: Proceedings of the International Conference on Computer Vision (ECCV), 2012, 525–538.
[7] M.F. Valstar, H. Gunes, M. Pantic, How to distinguish posed from spontaneous smiles using geometric features, in: Proceedings of the International Conference on Multimodal Interfaces, 2007, 38–45.
[8] P. Wu, H. Liu, X. Zhang, Spontaneous Versus Posed Smile Recognition Using Discriminative Local Spatio-temporal Descriptors, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2014, 1249–1253.
[9] 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.
[10] Y. Bai, L. Guo, L. Jin, Q. Huang, A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2009, 3305–3308.
[11] C. Shan, Smile detection by boosting pixel differences, IEEE Trans. Image Process. (TIP) 21 (1) (2012) 431–436.
[12] L. Wolf, T. Hassner, Y. Taigman, Similarity scores based on background samples, in: Proceedings of the Asian Conference of Computer Vision (ACCV), 2010, 88–97.
[13] E. Makinen, R. Raisamo, Evaluation of gender classification methods with automatically detected and aligned faces, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 30 (3) (2008) 541–547.
[14] I.A. Essa, A.P. Pentland, Coding, analysis, interpretation, and recognition of facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 19 (7) (1997) 757–763.
[15] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1) (1991) 71–86.
[16] 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.
[17] 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.
[18] L. Zhang, D. Tjondronegoro, V. Chandran, Random Gabor based templates for facial expression recognition in images with facial occlusion, Neurocomputing 145 (2014) 451–464.
[19] J. Whitehill, C.W. Omlin, Haar features for facs au recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FGR), 2006, 5–101.
[20] A. Hadid, M. Pietikainen, T. Ahonen, A discriminative feature space for detecting and recognizing faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2004, 797–804.
[21] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005, 886–893.
[22] 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.
[23] S. Walk, N. Majer, K. Schindler, B. Schiele, New features and insights for pedestrian detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, 1030–1037.
[24] M. Liu, S. Li, S. Shan, X. Chen, Au-inspired deep networks for facial expression feature learning, Neurocomputing 159 (2015) 126–136.
[25] A. Savran, B. Sankur, M.T. Bilge, Regression-based intensity estimation of facial action units, Image Vis. Comput. 30 (10) (2012) 774–784.
[26] J.R. Delannoy, J. McDonald, Automatic estimation of the dynamics of facial expression using a three-level model of intensity, in: Automatic Face & Gesture Recognition, 2008. FG'08, in: Proceedings of the 8th IEEE International Conference on, IEEE, 2008, 1–6.
[27] H. Dibeklioglu, A. Salah, T. Gevers, Recognition of genuine smiles, IEEE Trans. Multimed. 17 (3) (2015) 279–294.
[28] A. Miehlke, U. Fisch, C.-M. Eneroth, Surgery of the Facial Nerve, Saunders, 1973.
[29] P. Ekman, E.L. Rosenberg, What the Face Reveals: basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS), Oxford University Press, 1997.
[30] W.E. Rinn, The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions, Psychol. Bull. 95 (1) (1984) 52.
[31] J.F. Cohn, K.L. Schmidt, The timing of facial motion in posed and spontaneous smiles, Int. J. Wavel., Multiresolut. Inf. Process. 2 (2) (2004) 121–132.
[32] K.L. Schmidt, S. Bhattacharya, R. Denlinger, Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises, J. Nonverbal Behav. 33 (1) (2009) 35–45.
[33] K.L. Schmidt, Z. Ambadar, J.F. Cohn, L.I. Reed, Movement differences between

deliberate and spontaneous facial expressions: zygomaticus major action in smiling, J. Nonverbal Behav. 30 (1) (2006) 37–52.
[34] M. Hoque, R.W. Picard, Acted vs. natural frustration and delight: Many people smile in natural frustration, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG), 2011, 354–359.
[35] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.
[36] M.G. Frank, P. Ekman, Not all smiles are created equal: the differences between enjoyment and nonenjoyment smiles, Int. J. Humor Res. 6 (1) (1993) 9–26.
[37] E.G. Krumhuber, A.S. Manstead, Can Duchenne smiles be feigned? New evidence on felt and false smiles, Emotion 9 (6) (2009) 807.
[38] Z. Zhang, V. Singh, T.E. Slowe, S. Tulyakov, V. Govindaraju, Real-time automatic deceit detection from involuntary facial expressions, in: Proceedings of the Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, 1–6.
[39] T. Pfister, X. Li, G. Zhao, M. Pietikainen, Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework, in: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011, 868–875.
[40] H. Liu, P. Wu, Comparison of methods for smile deceit detection by training AU6 and AU12 simultaneously, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2012, 1805–1808.
[41] S.-J. Wang, W.-J. Yan, T. Sun, G. Zhao, X. Fu, Sparse Tensor Canonical Correlation Analysis for Micro-expression Recognition, Neurocomputing.
[42] V. Le, J. Brandt, Z. Lin, L. Bourdev, T. Huang, Interactive facial feature localization, in: Proceedings of the European Conference on Computer Vision (ECCV), 2012, 679–692.
[43] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, Int. J. Comput. Vis. (IJCV) 107 (2) (2014) 177–190.
[44] X.P. Burgos-Artizzu, P. Perona, P. Dollár, Robust face landmark estimation under occlusion, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2013, 1513–1520.
[45] P. Dollár, P. Welinder, P. Perona, Cascaded pose regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, 1078–1085.
[46] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 23 (6) (2001) 681–685.
[47] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.
[48] M. Hoque, L.-P. Morency, R.W. Picard, Are you friendly or just polite?–Analysis of smiles in spontaneous face-to-face interactions, in: Affective Computing and Intelligent Interaction, 2011, 135–144.
[49] H. Dibeklioglu, R. Valenti, A. A. Salah, T. Gevers, Eyes do not lie: spontaneous versus posed smiles, in: International Conference on Multimedia, 2010, 703–706.
[50] Y. Guo, G. Zhao, M. Pietikäinen, Discriminative features for texture description, Pattern Recognit. 45 (10) (2012) 3834–3843.
[51] Z. Guo, L. Zhang, D. Zhang, A completed modeling of local binary pattern operator for texture classification, IEEE Trans. Image Process. 19 (6) (2010) 1657–1663.
[52] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the mmi facial expression database, in: 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, 65, 2010.

Pingping Wu received a Ph.D. degree at the School of Electronics Engineering and Computer Science (EE & CS), Peking University (PKU), China, in 2016. Currently, she is working at the School of Technology, Nanjing Audit University, China. Her research interests are facial expression recognition, smile analysis, visual speech recognition. Related papers have been published on TMM, ICRA, ICIP, ICPR and ICASSP.

Hong Liu received a Ph.D. degree in mechanical electronics and automation in 1996, and serves as a Full Professor in the School of EE & CS, 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 Trans. on Signal Processing, and IEEE Trans. on PAMI.

Chao Xu received a Ph.D degree at the School of Computer Science, Wuhan University, China, in 2014. Currently, he serves as an associate professor in the School of Technology, Nanjing Audit University, China. His research interests include software reliability analysis, embedded systems and information system audit.



Yuan Gao received the B.E. degree in Intelligent Science and Technology from Xidian University in 2012. Then he obtained the M.S. degree in Computer Applied Technology from Peking University in 2015. From October of 2015, he studied in Kiel University of Germany as a PhD candidate student under the supervision of Prof. Dr. Reinhard Koch. His research interests include facial expression and gender recognition, camera calibration, dynamic light-field rendering.

Zheyuan Li received a B.E degree from University of Science and Technology Beijing in 2008. She is now a PhD candidate at the School of Electronics Engineering and Computer Science (EE & CS), Peking University (PKU), China. Her research interests lie in deep learning, object tracking, human action recognition and video surveillance.

Xuewu Zhang received a B.E degree in Communication Engineering from North China Electric Power University in 2013. He is working toward an M.S. degree at the School of Electronics Engineering and Computer Science (EE & CS), Peking University (PKU), Shenzhen, China. His research interests lie in facial expression recognition, smile analysis and video surveillance.