Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Spontaneous versus posed smile recognition via region-specific texture descriptor and geometric facial dynamics*

Ping-ping WU[1,2], Hong LIU‡[1,2], Xue-wu ZHANG[2], Yuan GAO[3]

([1]*MOE Key Laboratory of Machine Perception, Peking University, Beijing 100871, China*)

([2]*Engineering Lab on Intelligent Perception for Internet of Things, Peking University Shenzhen Graduate School, Shenzhen 518055, China*)

([3]*Department of Computer Science, Christian-Albrechts University, Kiel 24118, Germany*)

E-mail: pingpingwu@pku.edu.cn; hongliu@pku.edu.cn; zhangxuewu@sz.pku.edu.cn; yuan.gao@stu.uni-kiel.de

Received Jan. 27, 2016; Revision accepted May 22, 2016; Crosschecked June 3, 2017

Abstract: As a typical biometric cue with great diversities, smile is a fairly influential signal in social interaction, which reveals the emotional feeling and inner state of a person. Spontaneous and posed smiles initiated by different brain systems have differences in both morphology and dynamics. Distinguishing the two types of smiles remains challenging as discriminative subtle changes need to be captured, which are also uneasily observed by human eyes. Most previous related works about spontaneous versus posed smile recognition concentrate on extracting geometric features while appearance features are not fully used, leading to the loss of texture information. In this paper, we propose a region-specific texture descriptor to represent local pattern changes of different facial regions and compensate for limitations of geometric features. The temporal phase of each facial region is divided by calculating the intensity of the corresponding facial region rather than the intensity of only the mouth region. A mid-level fusion strategy of support vector machine is employed to combine the two feature types. Experimental results show that both our proposed appearance representation and its combination with geometry-based facial dynamics achieve favorable performances on four baseline databases: BBC, SPOS, MMI, and UvA-NEMO.

Key words: Facial landmark localization; Geometric feature; Appearance feature; Smile recognition
http://dx.doi.org/10.1631/FITEE.1600041 **CLC number:** TP37

1 Introduction

Smile is a valuable tool of nonverbal social communication, which can signal enjoyment, politeness or even cover other emotional feelings like embarrassment, tension, and fear (Ambadar *et al.*, 2009). Eighteen different types of smile were identified by Ekman (2009), who claimed that there may be as many as 50 types of smile. From the knowledge of neurology, the face is innervated by two different brain systems that compete for control of its muscles (Miehlke *et al.*, 1973). 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 mediated by these two systems show differences in both morphology and dynamics, which motivates us to use not only geometric but also appearance features. From the research in Rinn (1984) and

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 60675025), the National High-Tech R&D Program (863) of China (No. 2006AA04Z247), the Scientific and Technical Innovation Commission of Shenzhen Municipality, China (Nos. JCYJ20130331144631730 and JCYJ20130331144716089), and the Specialized Research Fund for the Doctoral Program of Higher Education, China (No. 20130001110011)

ORCID: Ping-ping WU, http://orcid.org/0000-0001-7822-5208

Ekman and Rosenberg (1997), facial expressions initiated by the sub-cortical system tend to be more symmetrical, consistent, and reflex-like, while facial expressions initiated by the cortical system tend to be less smooth and have more variable dynamics. Accordingly, all types of smile can be divided into two categories: voluntary (deliberated/fake/posed) and involuntary (spontaneous/genuine/felt). Specifically, Frank and Ekman (1993) narrated that only one particular type of smile, which accompanies with experienced positive emotions such as happiness, pleasure, and enjoyment, is called the enjoyment or spontaneous smile.

The significance of developing an automatic computer vision system to distinguish spontaneous and posed facial behaviors becomes apparent when one considers the different neurological substrates that mediate these two types of smile (Whitehill *et al.*, 2013). There are potential applications for spontaneous versus posed (SVP) smile recognition. People suffering from autism can use it in social interaction to detect deceptive facial expressions (Baron-Cohen *et al.*, 2000). Video cameras can employ this technique to capture not only a smile but also a natural and unforced smile. Human computer interaction (HCI) can be improved by deep understanding and interpretation of facial expressions.

As to SVP smile recognition, both spatial and temporal information needs to be extracted so that the recognition can be more accurate. This is different from tasks like face verification (Li *et al.*, 2005), gender classification (Liu *et al.*, 2014), and age estimation (Liu and Sun, 2016), which rely mainly on the spatial information and generally ignore the dynamic information in video sequences. Subtle smiles appear in a fleeting time as micro-expressions are not considered in the research scope of SVP smile recognition. For details about micro-expressions, refer to the specific research topic as micro-expression recognition (Pfister *et al.*, 2011b; Wu *et al.*, 2011; Shen *et al.*, 2012). There have been a few studies trying to distinguish spontaneous smiles from posed ones. Valstar *et al.* (2007) proposed a method to distinguish SVP smiles by fusing head, face, and shoulder modalities. Cohn and Schmidt (2004) observed that posed smiles are of larger amplitude and have a less consistent relationship between amplitude and duration than spontaneous ones. Hoque *et al.* (2012) explored temporal patterns to distinguish delighted smiles from frustrated smiles with their self-built database. Moreover, they found that acted instances are much easier for a computer to classify than natural ones. Dibeklioğlu *et al.* (2012) proposed a set of geometric features to distinguish genuine smiles from fake ones by calculating displacement signals of eyelids, cheeks, and lip corners. Besides, a corpus named UvA-NEMO was introduced, which includes 1240 video sequences and is the largest SVP smile database retrievable publicly. Recently, Dibeklioğlu *et al.* (2015) extended their work by adding experiments and analyses about effects of gender and age. Pfister *et al.* (2011a) proposed a spatiotemporal method to distinguish between spontaneous and posed facial expressions on a corpus including both natural and infrared videos. Specifically, they extended a spatial local descriptor to a spatiotemporal descriptor. In Liu and Wu (2012), a smile deceit detection was carried out by training AU6 and AU12 simultaneously on a static image database. Besides, an appearance-based local spatial-temporal descriptor was proposed in our recent work to distinguish between spontaneous and posed smiles using discriminative completed local binary pattern (CLBP) from three orthogonal planes (Wu *et al.*, 2014).

In summary, geometric features (GF) represent displacements of facial landmark, curvature changes of lips and eyelids, sizes of eyes, etc., through which facial dynamics such as amplitudes, speeds, and accelerations can be obtained. Appearance features characterizing texture information brought by facial muscle movements, like eye corner wrinkles, are an indispensable and non-substitutable element. The main contributions of this study are as follows:

1. It is the first time that both geometric and appearance features have been considered for SVP smile recognition to the best of our knowledge.

2. A region-specific texture descriptor is proposed to obtain discriminative and compact representations using only low-level features, where irregular facial regions are cropped with an advanced facial landmark localization method.

3. Taking into account the non-synchronous motion of different facial regions, temporal segmentation is implemented according to the corresponding facial region.

2 Temporal segmentation and geometric feature extraction

Geometric features rely on accurate and robust facial landmark localization. If facial landmarks can be precisely located and tracked, the accuracy of geometric feature extraction will be promoted. Therefore, to derive accurate geometric features, attention is paid to the state-of-the-art methods of facial landmark localization. Recently, great improvements have been made in this research field (Dollár *et al.*, 2010; Le *et al.*, 2012; Burgos-Artizzu *et al.*, 2013; Cao *et al.*, 2014), where facial landmarks can be located faster and more accurately than the traditional active appearance model (AAM) (Cootes *et al.*, 2001). Here, we employ a novel regression-based approach (Cao *et al.*, 2014) without using any parametric shape model, which has shown extraordinary performance in terms of both accuracy and efficiency. By using the method, up to 194 facial landmarks can be located accurately as shown in Fig. 1a with green dots. To be consistent with Dibeklioğlu *et al.* (2015), only 11 corresponding red landmarks are reserved as shown in Fig. 1a. Besides, three extra points are added, which are the mouth center (point 12) and the lower eyelid center (points 13 and 14). Note that all the reserved landmarks are shown in red and the generated landmarks like points 7, 8, and 12 are depicted in white (Fig. 1a). The generated landmarks are computed by the reserved landmarks.

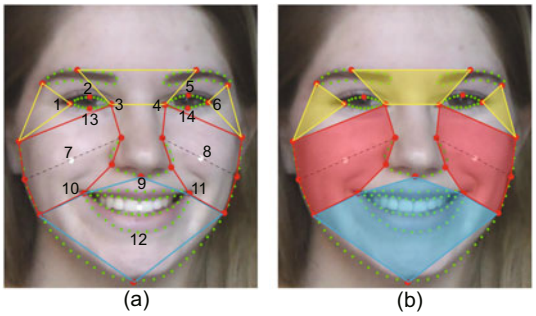

Fig. 1 Facial landmarks (a) and irregular facial region cropping (b) (References to color refer to the online version of this figure)

With respect to the 14 facial landmarks, face normalization is carried out. In Dibeklioğlu *et al.* (2015), smile temporal phases were segmented according only to the amplitude of lip corner movements. However, movements of lip corners (points 10 and 11), cheek centers (points 7 and 8), and eyelid centers (points 2, 5, 13, and 14) may not rise or decay simultaneously. Considering this problem, temporal segmentation is implemented independently for each facial region. Since abundant facial landmarks can be localized in each frame with high accuracies, the amplitude of lip corner movement is defined as

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

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

$$\mathcal{A}_{\text{eyelid}}(t)=\frac{d(p_2^t,p_{13}^t)+d(p_5^t,p_{14}^t)}{d(p_2^1,p_{13}^1)+d(p_5^1,p_{14}^1)}, \tag{2}$$

$$\mathcal{A}_{\text{cheek}}(t)=\frac{d(p_9^t,p_7^t)+d(p_9^t,p_8^t)}{d(p_9^1,p_7^1)+d(p_9^1,p_8^1)}. \tag{3}$$

The temporal segmentation for the cheek region is similar to that for the mouth region, where the longest continuous increase of $\mathcal{A}_{\text{cheek}}$ is defined as the rise phase, the longest decrease of $\mathcal{A}_{\text{cheek}}$ as the decay phase, and the sustain phase of the cheek region is between the last frame of the rise phase and the first frame of the decay phase. The eye region is different from both regions mentioned above, as the eye aperture becomes smaller with the increase of smile intensity. Note that eye blinks may accompany the smile process, affecting temporal segmentation of the eye region. By observing the amplitude computed using Eq. (2), it is found that blink lasts very short and the amplitude will contain sharp fluctuations when blinks occur. This is quite different from the relatively slow squinting process during the smile process. To eliminate the effect of eye blinks, eye blink detection is carried out by calculating the derivative of the eye amplitude. Then we compensate for the signal of blink areas and smooth the whole signal. Afterwards, the temporal segmentation for the eye region can be implemented by finding the longest continuous decrease of the smoothed signal as the rise phase and the longest increase as the decay phase. The sustain phase is between the last frame of the rise phase and the first frame of the

decay phase. Fig. 2 shows the amplitude and temporal segmentation for each facial region, verifying that the same temporal phase for the eye, cheek, and mouth starts and stops at different instances.

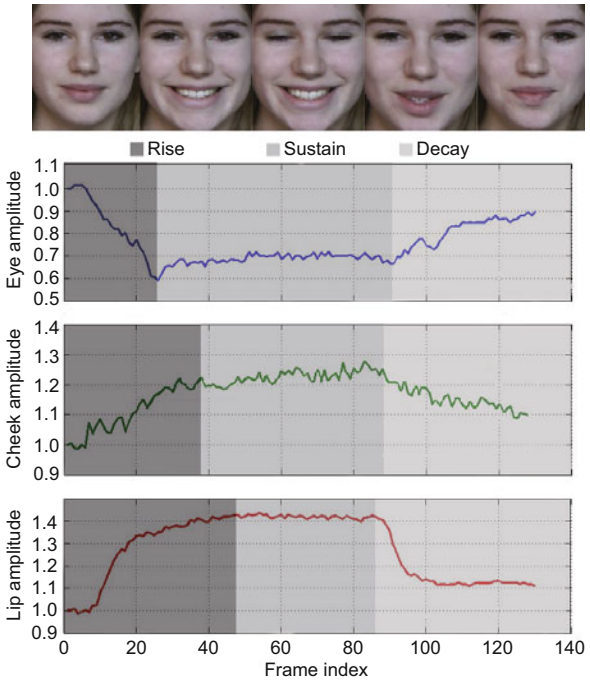

Fig. 2 Amplitudes and temporal phases in different facial regions

Sequentially, GFs representing facial dynamics can be extracted with the computed amplitude of each facial region. According to the calculation formulas proposed in Dibeklioğlu *et al.* (2015), as shown in Table 1, five types of geometric features are proposed which are amplitude related, amplitude versus duration related, duration related, speed related, and acceleration related. Speed $\mathcal{V}$ is the first derivative of amplitude $\mathcal{A}$ about time while acceleration $\mathcal{C}$ is the second derivative of amplitude $\mathcal{A}$ about time. Denote $\eta(\cdot)$ as the number of frames of a given signal. Duration is computed by dividing the number of frames occupied with frame rate ω. The defined GFs are extracted from each temporal phase separately. Therefore, three feature sets are obtained for each facial region. Additionally, in each temporal phase, increasing (+) and decreasing (−) are differentiated as the amplitude is not decreased or increased monotonically in each phase, as shown in Fig. 2. $\mathcal{A}_{\mathrm{L}}$ and $\mathcal{A}_{\mathrm{R}}$ indicate the amplitudes of the left and right sides of the face, respectively. However, compared to Dibeklioğlu *et al.* (2015), $\mathcal{A}_{\mathrm{eye}}$ and $\mathcal{A}_{\mathrm{cheek}}$ are calculated using our improved Eqs. (2) and (3) with an advanced method of facial landmark localization.

Table 1 Definitions of extracted geometric features

Feature	Definition
Duration	$\left[\frac{\eta(\mathcal{A}^+)}{\omega}, \frac{\eta(\mathcal{A}^-)}{\omega}, \frac{\eta(\mathcal{A})}{\omega}\right]$
Duration ratio	$\left[\frac{\eta(\mathcal{A}^+)}{\eta(\mathcal{A})}, \frac{\eta(\mathcal{A}^-)}{\eta(\mathcal{A})}\right]$
Maximum amplitude	$\max(\mathcal{A})$
Mean amplitude	$\left[\frac{\sum \mathcal{A}^+}{\eta(\mathcal{A}^+)}, \frac{\sum \lvert\mathcal{A}^-\rvert}{\eta(\mathcal{A}^-)}, \frac{\sum \mathcal{A}}{\eta(\mathcal{A})}\right]$
STD of amplitude	$\mathrm{std}(\mathcal{A})$
Total amplitude	$[\sum \mathcal{A}^+, \sum \lvert\mathcal{A}^-\rvert]$
Net amplitude	$[\sum \mathcal{A}^+ - \sum \lvert\mathcal{A}^-\rvert]$
Amplitude ratio	$\left[\frac{\sum \mathcal{A}^+}{\sum \mathcal{A}^+ + \sum \lvert\mathcal{A}^-\rvert}, \frac{\sum \lvert\mathcal{A}^-\rvert}{\sum \mathcal{A}^+ + \sum \lvert\mathcal{A}^-\rvert}\right]$
Maximum speed	$[\max(\mathcal{V}^+), \max(\lvert\mathcal{V}^-\rvert)]$
Mean speed	$\left[\frac{\sum \mathcal{V}^+}{\eta(\mathcal{V}^+)}, \frac{\sum \lvert\mathcal{V}^-\rvert}{\eta(\mathcal{V}^-)}\right]$
Maximum acceleration	$[\max(\mathcal{C}^+), \max(\lvert\mathcal{C}^-\rvert)]$
Mean acceleration	$\left[\frac{\sum \mathcal{C}^+}{\eta(\mathcal{C}^+)}, \frac{\sum \lvert\mathcal{C}^-\rvert}{\eta(\mathcal{C}^-)}\right]$
Net amplitude duration ratio	$\frac{(\sum \mathcal{A}^+ - \sum \lvert\mathcal{A}^-\rvert)\omega}{\eta(\mathcal{A})}$
Left/right amplitude difference	$\frac{\lvert\sum \mathcal{A}_{\mathrm{L}} - \sum \mathcal{A}_{\mathrm{R}}\rvert}{\eta(\mathcal{A})}$

3 Region-specific texture descriptor

3.1 Overview

According to the findings of psychology (Calvo and Nummenmaa, 2011; Calvo *et al.*, 2013), the smiling mouth is of higher visual saliency than eyes despite spontaneous or posed smiles. Therefore, the first fixation on a smiling mouth tends to confuse people to distinguish genuine smiles from fake ones, which leads to stored stereotyped knowledge about smiles rather than on the particular details of the actual smile. Consequently, a posed smile is more likely to be spotted when eye contact is made. Covering with the knowledge, it is reasonable and necessary to divide a smiling face into different regions to implement the recognition.

It can be easily observed that the extracted GF is not good at representing specific patterns like the wrinkle around outer eye corners and raised cheeks.

When using only GFs, a mouth-open non-smile facial image may be confused with a smile one in some cases. Moreover, in the recognition of six basic facial expressions, the surprise expression is of high possibility being confused with the happy expression. Therefore, designing an appearance feature may benefit the recognition of SVP smiles, which can make up for the underlying limitations of GFs to some extent. Specifically, a new process is designed to obtain a discriminative and compact representation for each facial region (Fig. 3). First, irregular facial regions (yellow, red, and blue) are cropped by the corresponding red facial landmarks as shown in Fig. 1b. Second, temporal segmentation for each facial region is carried out. Third, low-level features like raw pixel, Gabor, LBP, and histogram of oriented gradient (HOG) (Dalal and Triggs, 2005) are employed to extract texture features in each facial region and temporal phase as shown in Fig. 4. Then, locality-constrained linear coding (LLC) (Wang *et al.*, 2010) and temporal pooling are employed to improve the discriminability and compactness of low-level features, while whitened principal component analysis (PCA) is employed to narrow intra-class variations. Eventually, the final representation named region-specific texture descriptor (RSTD) is derived. Details about LLC, temporal pooling, and whitened PCA are presented in the following.

3.2 Locality-constrained linear coding

LLC was first proposed in Wang *et al.* (2010), which shows better performance than the common

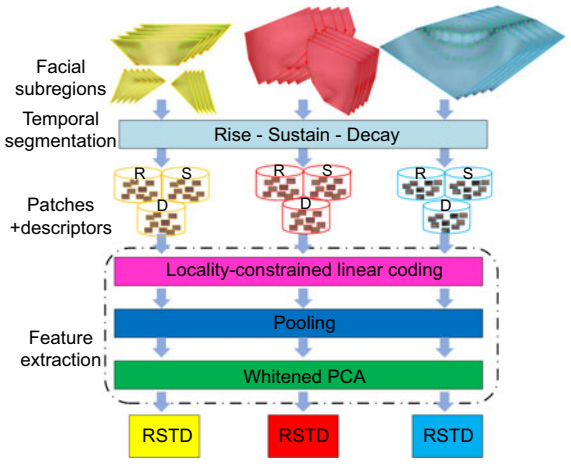

Fig. 3 Process of obtaining region-specific texture descriptors

Fig. 4 Patches where Pixel, Gabor, LBP, and HOG features are extracted (References to color refer to the online version of this figure)

coding schemes such as vector quantization and sparse coding. In this study, LLC is employed for its fast implementation and high efficiency (Wang *et al.*, 2012; Khokher *et al.*, 2014). Denote the over-complete sub-codebook as $\boldsymbol{B}_r^\tau$, which is learned from the rth facial region in the τth temporal phase $\tau \in \{1,2,3\}$ (1: rise phase; 2: sustain phase; 3: decay phase). The codebook $\boldsymbol{B}$ is constructed as follows:

$$\begin{aligned}\boldsymbol{B} &= \{\boldsymbol{B}_r^\tau | r = 1,2,\ldots,R, \tau = 1,2,3\},\\ \boldsymbol{B}_r^\tau &= [\boldsymbol{b}_{r,1}^\tau, \boldsymbol{b}_{r,2}^\tau, \ldots, \boldsymbol{b}_{r,M}^\tau] \in \mathbb{R}^{D\times M},\end{aligned} \tag{4}$$

where M is the number of entries in the codebook and $M \gg D$ (D is the dimension of the feature vector obtained), and $\boldsymbol{b}_{r,i}^\tau$, $i \in \{1,2,\ldots,M\}$, is the corresponding generated code vector. Sequentially, the region-specific feature $\boldsymbol{p}_i^\tau \in \mathbb{R}^D$, $i \in \{1,2,\ldots,N\}$, can be encoded using the following criterion:

$$\begin{aligned}&\min_{\boldsymbol{C}} \sum_{i=1}^{N} \|\boldsymbol{p}_i^\tau - \boldsymbol{B}_r^\tau \boldsymbol{c}_i\|^2 + \lambda \|\boldsymbol{d}_i^\tau \odot \boldsymbol{c}_i\|^2\\ &\text{s.t.} \quad \mathbf{1}^{\mathrm{T}} \boldsymbol{c}_i = 1, \forall i,\end{aligned} \tag{5}$$

where $\odot$ denotes the element-wise multiplication and $\boldsymbol{c}_i$ is the reconstructed vector. $\boldsymbol{d}_i^\tau \in \mathbb{R}^M$ is a locality adaptor which gives different freedom for each basis vector, and is proportional to the similarity of the input descriptor $\boldsymbol{p}_i^\tau$. It is computed as follows:

$$\boldsymbol{d}_i^\tau = \exp\left(\frac{\mathrm{dist}(\boldsymbol{p}_i^\tau, \boldsymbol{B}_r^\tau)}{\sigma}\right), \tag{6}$$

where

$$\begin{aligned}\mathrm{dist}(\boldsymbol{p}_i^\tau, \boldsymbol{B}_r^\tau) =& [\mathrm{dist}_{\mathrm{E}}(\boldsymbol{p}_i^\tau, \boldsymbol{b}_{r,1}^\tau), \mathrm{dist}_{\mathrm{E}}(\boldsymbol{p}_i^\tau, \boldsymbol{b}_{r,2}^\tau), \ldots,\\ & \mathrm{dist}_{\mathrm{E}}(\boldsymbol{p}_i^\tau, \boldsymbol{b}_{r,M}^\tau)],\end{aligned}$$

$\mathrm{dist}_{\mathrm{E}}(\boldsymbol{p}_i^\tau, \boldsymbol{b}_{r,j}^\tau)$ is the Euclidean distance between $\boldsymbol{p}_i^\tau$ and $\boldsymbol{b}_{r,j}^\tau$, and σ adjusts the weight decay speed of the locality adaptor.

3.3 Temporal pooling

As temporal phases have been segmented for each facial region, feature pooling is applied to the corresponding codes derived in each temporal phase. Since conventional max pooling can capture the salient properties (Yang *et al.*, 2009), improve the robustness, and make the subsequent work more convenient, it is employed here and computed as

$$\boldsymbol{x}_r^\tau = \max_{t\in[1,2,\ldots,T^\tau]} \max_{\boldsymbol{c}_i \in S_r^{\tau,t}} \boldsymbol{c}_i, \tag{7}$$

where T^τ is the number of frames in the τth temporal phase, $S_r^{\tau,t}$ is the tth frame from the rth region in the τth temporal phase, and $\boldsymbol{c}_i$ represents the ith code in $S_r^{\tau,t}$.

3.4 Whitened PCA

After the coding procedure, the derived feature vector $\boldsymbol{x}_r^\tau$ is of high dimension. Therefore, feature dimension needs to be reduced to obtain a compact representation. Considering that genuine smiles from different subjects have differences due to individual characteristics, whitened PCA (WPCA) is applied to suppress the difference from individual variations through the following steps: (1) Map feature $\boldsymbol{x}_r^\tau$ to the intra-class subspace by calculating the intra-class covariance matrix $\boldsymbol{C}_r^\tau \in \mathbb{R}^{M\times M}$ from the rth region in the τth temporal phase of all training examples. (2) Let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_g\}$ be the g largest eigenvalues of $\boldsymbol{C}_r^\tau$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_g]$ the corresponding eigenvectors. (3) Obtain a compact representation $\boldsymbol{a}_r^\tau$ for the rth region in the τth time phase as follows:

$$\boldsymbol{a}_r^\tau = \mathrm{diag}\left(\lambda_1^{-1/2}, \lambda_2^{-1/2}, \ldots, \lambda_g^{-1/2}\right) \boldsymbol{V}^{\mathrm{T}} \boldsymbol{x}_r^\tau. \tag{8}$$

Note that the inverses of the eigenvalues are multiplied with the features, which suppresses the responses from larger eigenvalues. Therefore, the difference from intra-class dissimilarity is reduced. The derived $\boldsymbol{a}_r^\tau$ is denoted as RSTD. At this point, the final feature representation including appearance features and facial dynamics introduced in Section 2 is achieved (Fig. 5).

4 Experiments and discussions

4.1 Setup

During the preprocessing of facial landmark regression, a rough face box is detected, and then the landmark is estimated in a coarse-to-fine way. Next geometric centers of the eyes and mouth can be detected. Then the whole face region is normalized using pre-defined ratio parameters for the whole video sequence. To verify RSTD, four conventional features—raw pixel, Gabor, HOG, and LBP—are employed as we aim at obtaining discriminative and compact features using RSTD with low-level features (Sariyanidi *et al.*, 2015). As the four features can be implemented only on rectangle-like images, how to use them in irregular facial regions becomes a problem. The cropped irregular facial regions are

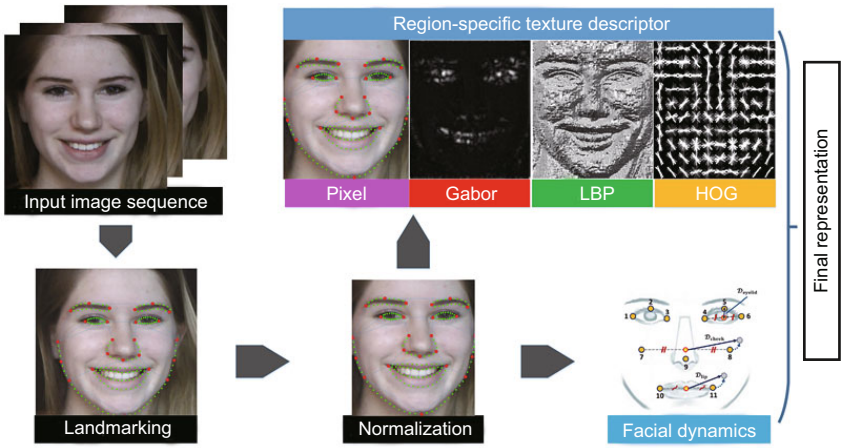

Fig. 5 Final feature representation including the appearance and geometric features, where the illustration of facial dynamics is quoted from Dibeklioğlu *et al.* (2015)

irregular polygons, which can be covered by overlapping or non-overlapping pathes. To derive abundant patch samples, overlapping patches are adopted as shown in Fig. 6. The detailed procedures to obtain patches from each irregular polygon are as follows: (1) Find the minimum enclosing rectangle (MER) of the irregular polygon. (2) Collect patches of $N \times N$ pixels with a stride of $k \times N$ pixels ($0 < k \leq 1$) in the obtained MER. (3) Retain the patches within the irregular polygon and discard the others not included in the polygon. The patch size is 16×16 pixels and k equals 0.5. For individual differences, the number of obtained patches is different for the same region. Taking the mouth region as an example, the mouth region sizes of different individuals are not the same, so different patch numbers are obtained. According to Fig. 4, we randomly choose 12 patches from the yellow irregular regions (eye regions), 50 patches from the red irregular regions (cheek regions), and 30 patches from the blue irregular region (mouth region). Afterwards, the features can be extracted and parameter settings of the four features are demonstrated as follows:

1. Raw pixel: the raw pixel feature is a feature vector constructed by all the pixel intensity values concatenating column by column in a patch, which is a simple baseline feature to refer to.

2. Gabor: Gabor filters are known for their similarity to the human biological visual system. There are two main parameters for the Gabor filter, orientation and scale. Here, eight orientations and five scales are chosen to form 40 Gabor filters, which are then downsampled by a factor of 4.

3. LBP: it is a typical local feature descriptor showing excellent performance in texture classification (Ojala *et al.*, 2002). Here, LBP$(8, 1, u_2)$ is applied, where u_2 indicates using a uniform pattern and $(8, 1)$ means sampling at 8 points with radius 1.

4. HOG: it is an appearance feature describing the shape of the object and has been successfully applied in pedestrian detection (Dalal and Triggs, 2005). HOG36 is employed in this study with a cell size of 4×4.

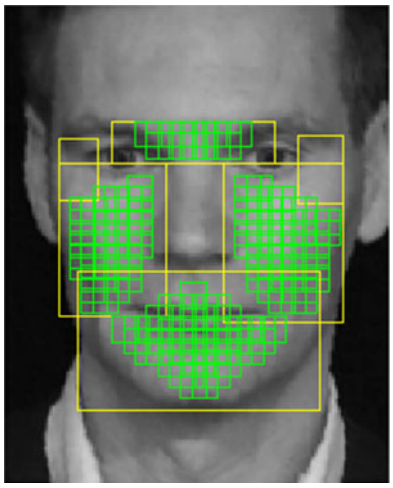

Fig. 6 An illustration of patches obtained in irregular facial regions. The patches in eye corner regions are not illustrated as the regions are much smaller than others (References to color refer to the online version of this figure)

4.2 Databases

To verify our proposed method for SVP smile recognition, several benchmark databases are employed, introduced briefly as follows:

1. UvA-NEMO database (http://www.e-nemo.nl) (Dibeklioğlu *et al.*, 2012) (the first two columns in Fig. 7) is so far the largest database for genuine and posed smile recognition. It includes 597 genuine smile videos and 643 posed ones collected at 50 frames/s 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.

2. SPOS corpus (http://www.ee.oulu.fi/gyzhao/) (Pfister *et al.*, 2011a) (the last two columns in Fig. 7) 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

Fig. 7 Exemplar frames from the UvA-NEMO (first and second columns) and SPOS (third and fourth columns) showing neutral face (first row), posed/deliberated smiles (second row), and the genuine/spontanesou smile (third row)

expression are employed. There are 66 genuine and 14 posed smiles captured with a resolution of 640 × 480 pixels at 25 frames/s in an indoor bunker environment involving 7 subjects (3 females and 4 males).

3. BBC database (Fig. 8) is from the 'Spot the fake smile' test on the BBC website (http://www.bbc.co.uk/science/humanbody/mind/surveys/smiles/), which consists of 10 genuine and 10 posed smile videos collected with a resolution of 314 × 286 pixels at 25 frames/s from 7 females and 13 males.

Fig. 8 All 20 subjects in the BBC database

4. MMI database (http://www.mmifacedb.eu) (Valstar and Pantic, 2010) is not specifically collected for SVP smile recognition. To compare with Dibeklioğlu *et al.* (2015), 74 posed smiles from 30 subjects are directly employed and 120 spontaneous smiles of 15 subjects are selected from 383 manually annotated segments. The database is recorded with two formats: 640 × 480 pixels at 29 frames/s and 720 × 576 pixels at 25 frames/s.

4.3 Evaluation on different features with different classification strategies

In this study, linear support vector machine (SVM) is applied for its simplicity, speed, and good performance. For feature fusion, there are three different fusion strategies of SVM, which are early, mid-level, and late fusions. 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 region, separately, and the region-based vectors are classified by SVMs; late fusion individually classifies features from each phase and region. For mid-level fusion and late fusion, a 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.

In Table 2, RSTD-Raw, RSTD-Gabor, RSTD-LBP, and RSTD-HOG denote that the proposed feature RSTD is trained and tested based on Raw, Gabor, LBP, and HOG, separately. The combination of all appearance features is denoted as RSTD-All. For GF, the robust facial landmark localization method introduced in Section 2 is used. To select discriminative features, the min-redundancy max-relevance (mRMR) algorithm (Peng *et al.*, 2005) is applied, the same as in Dibeklioğlu *et al.* (2015). Note that the GF in this study is a representation of geometric features with feature selection. The combination of geometric and appearance features is denoted as 'GF+RSTD-All'. Experiments here are implemented only on the UvA-NEMO database as it is the largest one with high resolution. We regard spontaneous smiles as positive samples while posed smiles as negative samples.

As shown in Table 2, the recognition rate of spontaneous smiles is generally a little bit lower than that of the posed ones, which may be caused by the fact that there are more examples of posed (negative) examples. Mid-level fusion performs the best for all features, followed by early and late fusions. The best performance of 93.95% is achieved by combining all feature types and using mid-level fusion. As to different appearance features, RSTD-Raw shows the worst performance while the performances of RSTD-Gabor and RSTD-HOG are better than that of RSTD-LBP. By combining all appearance features, RSTD-All improves the recognition accuracy compared to a single basic feature-based RSTD. However, we can also learn that there is great information redundancy among the appearance features. When using mid-level fusion, RSTD-All achieves an accuracy of 91.65% while GF achieves 89.06%. It can be seen that the appearance feature-based RSTD-All is slightly superior to GF, while their combination achieves the best performance. With regard to GF,

Table 2 Comparison of different features using different classification strategies on UvA-NEMO

Strategy	True positive rate (%)			True negative rate (%)			Hit rate (%)		
	Early fusion	Mid-level	Late fusion	Early fusion	Mid-level	Late fusion	Early fusion	Mid-level	Late fusion
RSTD-Raw	59.63	60.65	55.78	60.65	64.07	59.72	60.17	62.35	57.84
RSTD-Gabor	80.23	83.91	78.39	82.58	85.85	81.18	81.40	84.43	79.82
RSTD-LBP	73.87	75.21	72.19	75.89	79.16	76.67	74.93	77.21	74.52
RSTD-HOG	78.06	81.57	76.55	81.18	83.83	78.07	79.65	82.76	77.34
RSTD-All	87.10	90.62	85.43	90.67	92.53	87.40	88.90	**91.65**	86.48
GF	85.92	86.59	82.24	88.65	91.29	89.27	87.31	**89.06**	85.92
GF+RSTD-All	90.95	93.13	88.78	92.69	94.71	92.07	91.87	**93.95**	90.51

the combination improves the recognition accuracy by about 5%, which shows that the appearance feature is with the ability to compensate for the geometric feature. To the best of our knowledge, it is the first time that both appearance and geometric features have been considered for the task of SVP smile recognition while the state-of-the-art facial landmark localization method is used, aiming at deriving more precise geometric features.

4.4 Evaluation on different temporal phases and facial regions

In Section 2, inspired by the knowledge from psychology, the facial region is divided into several irregular regions and the amplitude of each facial region is defined. Then, the temporal segmentation for each facial region is carried out according to the corresponding defined amplitude. This is different from the work in Dibeklioğlu *et al.* (2012; 2015), which divides the temporal phases of each facial region using only the amplitude of lip corner movement. For a fair comparison, we use the same type of feature, i.e., GF, and perform the same feature selection procedure. Table 3 shows the recognition accuracies on UvA-NEMO in different facial regions and temporal phases using different temporal segmentation methods. As the geometric feature in our study is based on different facial landmark localization methods compared to Dibeklioğlu *et al.* (2012; 2015), the overall results are improved due to the high accuracy of the localized landmarks. The final result of GF is 86.37% (quoted from Dibeklioğlu *et al.* (2015)) while ours is 89.06%.

As to 'All phases' in Table 3, feature vectors from all three phases concatenate into one vector as the input for the SVM classifier; for 'All regions', feature vectors from all facial regions concatenate into one as the input for the SVM classifier. Compared with the temporal segmentation in Dibeklioğlu *et al.* (2015), the classification accuracies get improved especially for the eye region. As the temporal segmentations of the two methods are the same for the lip region, the accuracies in all three phases are identical. This verifies that the temporal phase segmentation for the eye is different from that for the lip. In fact, during most of the time, movements in the two regions are not simultaneous. For cheek regions, the rise phase and decay phase achieve some slight improvement using our segmentation method, while the sustain phase shows a 0.07% recognition accuracy reduction. This may be caused by inaccuracy of landmarks 7 and 8 (see Fig. 1), as they are not directly detected using landmark localization but computed using other detected landmarks. The rise

Table 3 Recognition accuracies on UvA-NEMO in different facial regions and temporal phases using different temporal segmentation methods

Facial region	Recognition accuracy (%)			
	Onset, Rise	Apex, Sustain	Offset, Decay	All phases
Eye	81.04, **84.28**	72.31, **73.54**	65.16, **67.90**	87.15, **88.04**
Cheek	78.68, 79.34	69.39, 69.32	61.17, 62.25	83.62, 83.81
Lip	83.75, 83.75	70.76, 70.76	58.93, 58.93	85.85, 85.85
All regions	85.51, **87.69**	72.07, 75.46	68.20, 70.78	87.63, **89.06**

Data on the left of the comma show results using the temporal segmentation method in Dibeklioğlu *et al.* (2015), while data on the right of the comma show results using our temporal segmentation method

phase shows the most discriminant power for all facial regions, followed by sustain and decay phases. The accuracies in the decay phase are not so ideal, which is consistent with the fact that when a smile comes to the decay phase, its intensity may show large fluctuations accompanying subordinate smiles. Using our segmentation method, the eye region is the most reliable region in all phases, followed by the lip and cheek regions. The eye region shows more discriminative power than the lip region for all temporal phases. An accuracy of 84.28% is acquired for the eye region in the rise phase, while an approximative accuracy of 83.75% is achieved for the lip region in this phase. This is different from the result in Dibeklioğlu *et al.* (2015), where the lip region in the rise phase achieves a better performance than the eye region. The reason giving rise to this difference is due mainly to our new definition of the calculation of the eye amplitude.

Since a face is divided into three regions, experiments are implemented on each facial region in each temporal phase. In this way, the discrimination power of each region and phase using RSTD-All and its combination with geometric features can be investigated (Fig. 9).

Obviously, the appearance feature RSTD-All is more discriminative than GF for the eye region in all temporal phases. As shown in Fig. 9, the performance of RSTD-All on the eye region in the rise phase is close to 90% while the performance of GF is lower than 85%. However, it shows the opposite result for the mouth region, where GF achieves an accuracy of 83.75% in the rise phase while RSTD-All achieves an accuracy of 80.56%. Using geometric features, the mouth region shows more discriminative power, which is consistent with the result in Dibeklioğlu *et al.* (2012) using only geometric features. Alternatively, using the appearance feature, the eye region achieves the best performance. From the psychology research in Calvo *et al.* (2013), one of the main differences between genuine and fake smiles is the movement of eyes. Appearance features are able to capture the significant patterns around eyes like eye crow's feet while geometric features have little such ability. For the mouth region, geometric facial dynamics plays a more important role than appearance features. Based on the above analyses, it may come to a conclusion that there is some texture/appearance difference in the eye region for the

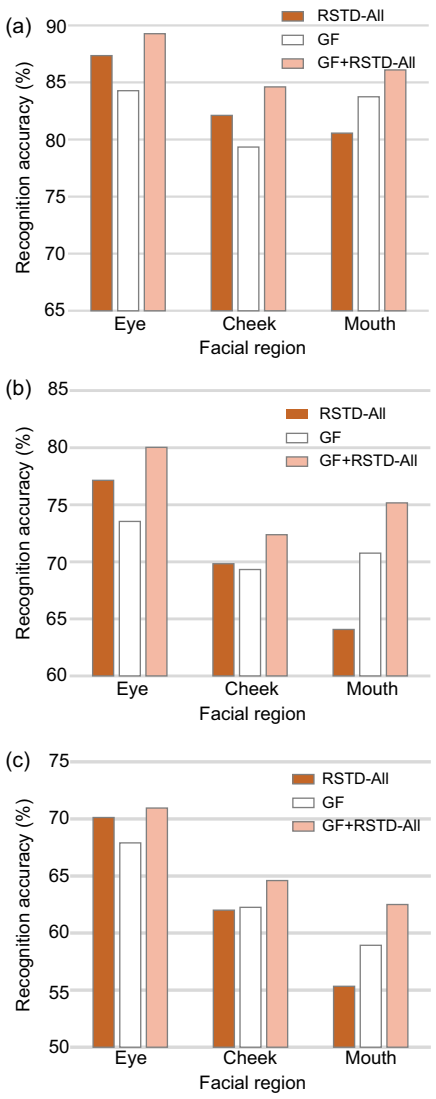

Fig. 9 Assessment for appearance and geometric features in different temporal phases and facial regions: (a) rise phase; (b) sustain phase; (c) decay phase

two types of smile, which can be better characterized by the appearance feature. Back to Fig. 9a, the best performance is achieved by the combination of two features on the eye region followed by the mouth and cheek regions. As shown in Figs. 9b and 9c, in sustain and decay phases, both features on the mouth region show unfavorable performances, especially for the appearance feature RSTD-All. However, after combining the two features on this region, a notable improvement is achieved, which again demonstrates that the two feature types can complement each other. From Figs. 9b and 9c, we can also observe that the performances of the two feature types are similar on the cheek region. However, from Fig. 9a,

the appearance feature plays a more effective role on the cheek region. Attention should be paid to the scale of the vertical coordinates. Note that the best performance is achieved in the rise phase, followed by the sustain and decay phases. Also, note that the computation complexity of the proposed appearance feature RSTD is low as it is linear to the size of the codebook and the number of sampled pathes.

4.5 Comparison with other methods

Comparisons of our method with the state-of-the-art SVP smile recognition methods are carried out on the four databases introduced previously. The infrared images of SPOS are not used since the other three datasets do not contain such content. The recognition results of Cohn and Schmidt (2004), Pfister *et al.* (2011a), and Dibeklioğlu *et al.* (2012; 2015) are directly quoted from Dibeklioğlu *et al.* (2012; 2015). For our method, the proposed appearance feature RSTD-All and its combination with geometric facial dynamics GF are both implemented. The other experimental protocols are the same, employing a two-level 10-fold cross validation scheme detailedly presented in Dibeklioğlu *et al.* (2015). Furthermore, SPOS involves only seven subjects, and one subject is used for the database.

As shown in Table 4, the fusion of RSTD-ALL and GF outperforms the others on UvA-NEMO, SPOS, and MMI, achieving the highest performance of 93.95%, 81.25%, and 92.21%, respectively. Besides, RSTD-All shows competitive performance compared to other methods. It outperforms the methods on all datasets but SPOS, where its recognition accuracy is 0.25% lower than that in Wu *et al.* (2014). The recognition accuracy remains 90.00% on BBC, which is the same as those in Dibeklioğlu *et al.* (2012), Wu *et al.* (2014), and Dibeklioğlu *et al.* (2015). The recognition rates on SPOS of all methods are lower than those on three other databases. One of the main reasons is that BBC and SPOS are of low resolution, which makes appearance features insufficient or lost to some extent. Moreover, SPOS contains only the onset phase of smiles while the two other phases are not captured, resulting in the loss of dynamic information. What is more, Dibeklioğlu *et al.* (2015) is an expansion of Dibeklioğlu *et al.* (2012), where the recognition accuracy is improved from 87.02% to 92.90% by adding age group information in the feature. Naturally, with the growth of age, the appearance and texture of human face change. Therefore, adding age group features can be regarded as adding appearance information, which is verified by the experimental results that combining the geometric feature with the appearance feature results in the best accuracy.

5 Conclusions

Motivated by the visual saliency of smiling faces and the limitation of geometry-based features, we propose a novel appearance representation—region-specific texture descriptor (RSTD)—based on irregular facial regions, to capture discriminative and robust texture information. Temporal segmentation based on each facial region is presented, which is more sensible than the segmentation by using only the amplitude of the lip region. Using both appearance and geometric features achieves better performance than using either one of them, which demonstrates that the two types of information are both indispensable for spontaneous versus posed smile recognition. Furthermore, experiments on four benchmark databases show that our proposed appearance representation RSTD is rather competitive,

Table 4 Correct recognition rates on four public databases

Method	Recognition accuracy (%)			
	UvA-NEMO	BBC	SPOS	MMI
RSTD-All+GF (Ours)	**93.95**	**90.00**	**81.25**	**92.21**
RSTD-All (Ours)	91.65	90.00	79.25	90.72
Dibeklioğlu *et al.* (2015)	92.90	90.00	78.75	90.21
Wu *et al.* (2014)	91.40	90.00	79.50	86.10
Dibeklioğlu *et al.* (2012)	87.02	90.00	75.00	86.43
Dibeklioğlu *et al.* (2010)	71.05	85.00	66.25	72.55
Pfister *et al.* (2011a)	73.06	70.00	67.50	81.37
Cohn and Schmidt (2004)	77.26	75.00	72.50	79.02

verifying the effectiveness of our proposed method. Moreover, its fusion with geometric facial dynamics outperforms the state-of-the-art methods, which shows that the two feature types can complement for each other. Finally, we believe our work can be applied as a tool for the analysis of smiles in psychology and help people suffering from autism well interpret the expression. It can also provide a better and deeper way for human robot interaction. In our future work, temporal-order preserving coding methods will be explored to cooperate with spatial configuration coding methods like locality-constrained linear coding.

References

Ambadar, Z., Cohn, J.F., Ian Reed, L., 2009. All smiles are not created equal: morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *J. Nonverb. Behav.*, **33**(1):17-34. https://doi.org/10.1007/s10919-008-0059-5

Baron-Cohen, S., Ring, H.A., Bullmore, E.T., *et al.*, 2000. The amygdala theory of autism. *Neurosci. Biobehav. Rev.*, **24**(3):355-364. https://doi.org/10.1016/S0149-7634(00)00011-7

Burgos-Artizzu, X.P., Perona, P., Dollár, P., 2013. Robust face landmark estimation under occlusion. Proc. IEEE Int. Conf. on Computer Vision, p.1513-1520. https://doi.org/10.1109/ICCV.2013.191

Calvo, M.G., Nummenmaa, L., 2011. Time course of discrimination between emotional facial expressions: the role of visual saliency. *Vis. Res.*, **51**(15):1751-1759. https://doi.org/10.1016/j.visres.2011.06.001

Calvo, M.G., Gutiérrez-García, A., Avero, P., *et al.*, 2013. Attentional mechanisms in judging genuine and fake smiles: eye-movement patterns. *Emotion*, **13**(4):792-802. http://dx.doi.org/10.1037/a0032317

Cao, X., Wei, Y., Wen, F., *et al.*, 2014. Face alignment by explicit shape regression. *Int. J. Comput. Vis.*, **107**(2):177-190. https://doi.org/10.1007/s11263-013-0667-3

Cohn, J.F., Schmidt, K.L., 2004. The timing of facial motion in posed and spontaneous smiles. *Int. J. Wavel. Multiresol. Inform. Process.*, **2**(2):121-132. https://doi.org/10.1142/S021969130400041X

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. *IEEE Trans. Patt. Anal. Mach. Intell.*, **23**(6):681-685. https://doi.org/10.1109/34.927467

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.886-893. https://doi.org/10.1109/CVPR.2005.177

Dibeklioğlu, H., Valenti, R., Salah, A.A., *et al.*, 2010. Eyes do not lie: spontaneous versus posed smiles. Proc. Int. Conf. on Multimedia, p.703-706. https://doi.org/10.1145/1873951.1874056

Dibeklioğlu, H., Salah, A., Gevers, T., 2012. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. Proc. European Conf. on Computer Vision, p.525-538. https://doi.org/10.1007/978-3-642-33712-3_38

Dibeklioğlu, H., Salah, A., Gevers, T., 2015. Recognition of genuine smiles. *IEEE Trans. Multimed.*, **17**(3):279-294. https://doi.org/10.1109/TMM.2015.2394777

Dollár, P., Welinder, P., Perona, P., 2010. Cascaded pose regression. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1078-1085. https://doi.org/10.1109/CVPR.2010.5540094

Ekman, P., 2009. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. W. W. Norton & Company, New York, p.140-143.

Ekman, P., Rosenberg, E.L., 1997. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press.

Frank, M.G., Ekman, P., 1993. Not all smiles are created equal: the differences between enjoyment and nonenjoyment smiles. *Humor*, **6**(1):9-26. https://doi.org/10.1515/humr.1993.6.1.9

Hoque, M., McDuff, D., Picard, R., 2012. Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Trans. Affect. Comput.*, **3**(3):323-334. https://doi.org/10.1109/T-AFFC.2012.11

Khokher, M.R., Bouzerdoum, A., Phung, S.L., 2014. Crowd behavior recognition using dense trajectories. Proc. Int. Conf. on Digital lmage Computing: Techniques and Applications, p.1-7. https://doi.org/10.1109/DICTA.2014.7008098

Le, V., Brandt, J., Lin, Z., *et al.*, 2012. Interactive facial feature localization. Proc. European Conf. on Computer Vision, p.679-692. https://doi.org/10.1007/978-3-642-33712-3_49

Li, W.S., Zhou, C.L., Xu, J.T., 2005. A novel face recognition method with feature combination. *J. Zhejiang Univ.-Sci.*, **6A**(5):454-459. https://doi.org/10.1631/jzus.2005.A0454

Liu, H., Sun, X., 2016. A partial least squares based ranker for fast and accurate age estimation. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.2792-2796. https://doi.org/10.1109/ICASSP.2016.7472186

Liu, H., Wu, P., 2012. Comparison of methods for smile deceit detection by training AU6 and AU12 simultaneously. Proc. IEEE Int. Conf. on Image Processing, p.1805-1808. https://doi.org/10.1109/ICIP.2012.6467232

Liu, H., Gao, Y., Wang, C., 2014. Gender identification in unconstrained scenarios using self-similarity of gradients features. Proc. IEEE Int. Conf. on Image Processing, p.5911-5915. https://doi.org/10.1109/ICIP.2014.7026194

Miehlke, A., Fisch, U., Eneroth, C.M., 1973. Surgery of the Facial Nerve. Saunders, Philadelphia.

Ojala, T., Pietikäinen, M., Mäenpää, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Patt. Anal. Mach. Intell.*, **24**(7):971-987. https://doi.org/10.1109/TPAMI.2002.1017623

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Patt. Anal. Mach. Intell.*, **27**(8):1226-1238. https://doi.org/10.1109/TPAMI.2005.159

Pfister, T., Li, X., Zhao, G., *et al.*, 2011a. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. Proc. IEEE Int. Conf. on Computer Vision Workshops, p.868-875. https://doi.org/10.1109/ICCVW.2011.6130343

Pfister, T., Li, X., Zhao, G., *et al.*, 2011b. Recognising spontaneous facial micro-expressions. Proc. IEEE Int. Conf. on Computer Vision, p.1449-1456. https://doi.org/10.1109/ICCV.2011.6126401

Rinn, W.E., 1984. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychol. Bull.*, **95**(1):52-77. https://doi.org/10.1037/0033-2909.95.1.52

Sariyanidi, E., Gunes, H., Cavallaro, A., 2015. Automatic analysis of facial affect: a survey of registration, representation, and recognition. *IEEE Trans. Patt. Anal. Mach. Intell.*, **37**(6):1113-1133. https://doi.org/10.1109/TPAMI.2014.2366127

Shen, X.B., Wu, Q., Fu, X.L., 2012. Effects of the duration of expressions on the recognition of microexpressions. *J. Zhejiang Univ.-Sci. B*, **13**(3):221-230. http://dx.doi.org/10.1631/jzus.B1100063

Valstar, M., Pantic, M., 2010. Induced disgust, happiness and surprise: an addition to the MMI facial expression database. Proc. 3rd Int. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, p.65-70.

Valstar, M.F., Gunes, H., Pantic, M., 2007. How to distinguish posed from spontaneous smiles using geometric features. Proc. Int. Conf. on Multimodal Interfaces, p.38-45. https://doi.org/10.1145/1322192.1322202

Wang, J., Yang, J., Yu, K., *et al.*, 2010. Locality-constrained linear coding for image classification. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.3360-3367. https://doi.org/10.1109/CVPR.2010.5540018

Wang, X., Wang, L., Qiao, Y., 2012. A comparative study of encoding, pooling and normalization methods for action recognition. Proc. Asian Conf. on Computer Vision, p.572-585. https://doi.org/10.1007/978-3-642-37431-9_44

Whitehill, J., Bartlett, M.S., Movellan, J.R., 2013. Automatic facial expression recognition. *In*: Gratch, J., Marsella, S. (Eds.), Social Emotions in Nature Artifact. Oxford Scholarship Online. https://doi.org/10.1093/acprof:oso/9780195387643.003.0007

Wu, P.P., Liu, H., Zhang, X.W., 2014. Spontaneous versus posed smile recognition using discriminative local spatio-temporal descriptors. Proc. Int. IEEE Conf. on Acoustics, Speech and Signal Processing, p.1249-1253. https://doi.org/10.1109/ICASSP.2014.6853795

Wu, Q., Shen, X.B., Fu, X.L., 2011. The machine knows what you are hiding: an automatic micro-expression recognition system. *LNCS*, **6975**:152-162. https://doi.org/10.1007/978-3-642-24571-8_16

Yang, J.C., Yu, K., Gong, Y.H., *et al.*, 2009. Linear spatial pyramid matching using sparse coding for image classification. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1794-1801. https://doi.org/10.1109/CVPR.2009.5206757