SPONTANEOUS VERSUS POSED SMILE RECOGNITION USING DISCRIMINATIVE LOCAL SPATIAL-TEMPORAL DESCRIPTORS

Pingping Wu, Hong Liu, Xuewu Zhang

Engineering Lab on Intelligent Perception for Internet of Things (ELIP)
Key Laboratory of Machine Perception and Intelligence
Shenzhen Graduate School, Peking University, China
Email: pingpingwu@pku.edu.cn, hongliu@pku.edu.cn, zhangxuewu@sz.pku.edu.cn

ABSTRACT

Automatic recognition of spontaneous versus posed (SVP) facial expressions has received widespread attention in recent years for its potential applications in friendly human machine interface. Most existing works of SVP facial expression recognition extract geometry-based features which heavily rely on accurate detection and tracking of facial feature points. In this paper, a novel approach is proposed to distinguish between spontaneous and posed smiles using discriminative completed LBP from three orthogonal planes, which is an appearance-based local spatial-temporal descriptor. The descriptor devotes to extracting most robust and discriminative patterns of interest. In addition, flexible facial subregion cropping, a spatial division method, is proposed taking into account different facial organ size of different people and filtering of redundant information. Besides, in the temporal domain, a new division method is also applied, which divides the smile process according to smile dynamics . Experiments on three benchmark databases and comparisons to the state-of-the-art methods validate the advantages of our approach, obtaining an accuracy rate of 91.40% .

Index Terms— Smile Recognition, Spontaneous versus Posed, LBP

1. INTRODUCTION

Not all smiles are created equal as the diversity of smiles. However, studies of cognitive science indicate that all smiles can be divided into two categories: spontaneous and posed smiles [1]. Automatic spontaneous versus posed (SVP) smile recognition is necessary for visual analysis of social interaction signals. There are substantial potential applications for SVP smile recognition. People suffering from autism [2] can use SVP in social interaction to detect deceptive facial expressions. Video cameras can employ this technique to capture not only a smile but also a natural and unforced smile.

Facial expression recognition has been an active research area over the last decade [3, 4]. For recent years, SVP facial expression recognition has also gained a lot attention while many related literatures are published [5–9]. M. Valstar et al. proposed a method to distinguish SVP smiles by fusing head, face, and shoulder modalities [10]. E. Hoque et al. explored temporal patterns to distinguish delighted smiles from frustrated smiles using a facial feature tracker [11]. H. Dibeklioğlu et al. proposed an approach to spot SVP smiles using geometry-based features and the largest SVP smile database named UvA-NEMO was collected by them to date [12]. Pfister et al. utilized a local texture descriptor to distinguish between SVP facial expressions [13]. In our earlier work [14], a smile deceit detection has been done by training AU6 and AU12 simultaneously on a static-image database.

Generally, most proposed methods for SVP smile recognition extract geometry-based features [10–12]. However, geometry-based features commonly rely on accurate detection and tracking of facial fiducial points. In this paper, we improve an appearance-based feature proposed in [13] to distinguish SVP smiles instead of using geometry-based features. The main contributions of this paper are as follows: 1) discriminative completed LBP from three orthogonal planes (disCLBP-TOP) is proposed which is capable of extracting robust and discriminative features; 2) a new division method in spatial and temporal domain are respectively brought forward to enhance the performance of disCLBP-TOP.

2. DIVISION IN SPATIAL-TEMPORAL DOMAIN

2.1. Flexible Facial Subregion Cropping

Since disCLBP-TOP is a local descriptor, global information is absent. To overcome the defect, the whole image sequence is usually equally divided into blocks in both spatial and temporal domain. Unlike the equal division, flexible facial subregion cropping (FSC) is proposed considering specific facial regions for previous studies have shown that the subregions such as eyes play important roles for SVP smile recogni-

This work is supported by National Natural Science Foundation of China (NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No.JCYJ20120614152234873, CXC201104210010A).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tion [15, 16]. Besides, it also takes into account that different subjects' facial organs are of different size and changeable when speech and expressions occur. Five points (center of eyes s_1,s_2, lip corners s_3,s_4, nose tip s_5) are detected and tracked to locate facial subregions as shown in Fig. 1(a). Facial subregion volumes illustrated in Fig. 1 (b) can be derived with Algorithm 1, where α_1, α_2, β_1, and β_2 are the parameters controlling region size according to the prior knowledge of face proportion.

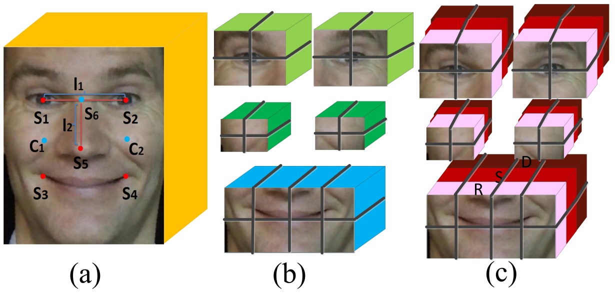

Fig. 1: (a) Facial key points of a subject from UvA-NEMO (b) Cropped subregion volumes (c) Divided blocks in spatial-temporal domain

Algorithm 1: FSC

```
   Input: a smile sequence $\{I_i|i = 1, ..., k\}$, $k$ is the
          number of frames;
   Output: subvolumes $V$
1  for each frame do
2      detect and track $\{s_m|m = 1, ..., 5\}$;
3      compute midpoint $s_6$ of $s_1$ and $s_2$, $c_1$ of $s_1$ and $s_3$
       and $c_2$ of $s_2$ and $s_4$;
4      compute distance $l_1$ = $|s_1$ - $s_2|$ , $l_2$ = $|s_6$ -$s_5|$;
5      crop the left and right eye region $R_1$, $R_2$ with
       central point $s_1$, $s_2$, width $\alpha_1 l_1$, height $\alpha_2 l_2$
       respectively ;
6      crop the left and right cheek region $R_3$, $R_4$ with
       central point $c_1$, $c_2$, width $\beta_1 l_1$, height $\beta_2 l_2$
       respectively ;
7      locate the mouth region $R_5$ with $s_3$,$s_4$ and $s_5$;
8  end
9  form region volumes $V_j$ with normalized $R_j^i$,
   $V_j = \{R_j^i|i = 1, ..., k\}, j = 1, ..., 5$;
10 for $j = 1$ to $4$ do
11     divide $V_j$ into 2× 2 subvolumes $\{V_{j,r}|r = 1, ..., 4\}$
12 end
13 if $j = 5$ then
14     divide $V_j$ into 2× 4 subvolumes $\{V_{j,t}|t = 1, ..., 8\}$
15 end
16 return $V$ = $\{V_{j,r}|r = 1, ..., 4\} \bigcup \{V_{j,t}|t = 1, ..., 8\}$
```

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 size; 3) some redundant information could be filtered out such as nose and forehead which are relatively static. All the advantages are favorable for the statistics of disCLBP-TOP. However, over-division in subvolumes (step 10 to 15 in Algorithm 1) may make a similar effect as a direct uniform division of face area.

2.2. Time Division

In temporal domain, a smile process is usually divided into three phases (onset, apex and offset) as applied in [17]. Another three phases (rise, sustain and decay) are applied here to better analyze smile dynamics, which is different from the work in [18] dividing time axis into equal length. As shown in Fig. 2(b), smiles often have a sustained region with multiple peaks, thus there is often not one clear apex or peak to the smile. The smile detector from OpenCV is improved and employed here. The normalized smile intensity I_{smile}^i in the ith frame is estimated in each frame as:

$$I_{smile}^i = \frac{S_n^i - N}{M - N + 1} \tag{1}$$

where S_n^i denotes the number of current detected smile neighbors while M and N are the maximum and minimum number of detected smile neighbors respectively. An intensity threshold θ is predetermined to mark the start or end of the sustain period. Through this way, each subvolume V can be divided into three blocks as shown in Fig. 1(c).

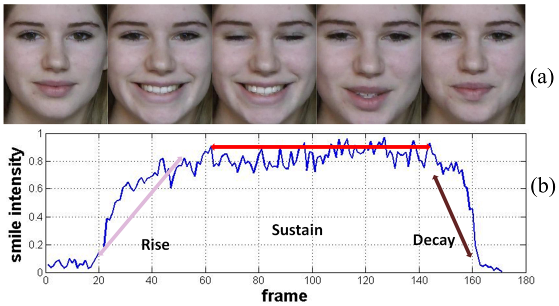

Fig. 2: (a) a deliberate smile sequence from UvA-NEMOA (b) visual example of a smile's Rise, Decay and Sustain

3. DISCRIMINATIVE LOCAL SPATIAL-TEMPORAL DESCRIPTOR: DISCLBP-TOP

Completed local binary pattern (CLBP) extends local binary pattern (LBP) by adding the local difference of its central pixel intensity (C) and magnitude (M) besides sign (S), which has shown good performance in texture classification [19]. In [13], in order to derive dynamic information, the purely spatial CLBP was first extended to spatial-temporal domain by extracting CLBP features from three orthogonal planes (CLBP-TOP). If applying CLBP-TOP directly to each divided block and then concatenating the histograms, the feature

Algorithm 2: Learning process of disCLBP-TOP with FSC and time division

```
Input: class $c$ with $n_c$ examples $\{S_1, S_2, ..., S_{n_c}\}$, B is the number of blocks
Output: extracted feature $J_{Global}$ of class $c$
for $n = 1$ to $n_c$ do
    divide $S_n$ into blocks $\{B_v|v = 1, ..., B\}$ with FSC and time division;
    compute sign dominant pattern $JS^n_{u,v}$ in $B_v$, $u = 1, 2, 3$;
    compute magnitude dominant pattern $JM^n_{u,v}$ in $B_v$, $u = 1, 2, 3$;
    $JS_{u,v} = JS_{u,v} \bigcap JS^n_{u,v}$;
    $JM_{u,v} = JM_{u,v} \bigcap JM^n_{u,v}$;
end
for $v = 1$ to $B$ and $u = 1$ to $3$ do
    $JS_{Global}$ = $JS_{Global} \bigcup JS_{u,v}$;
    $JM_{Global}$ = $JM_{Global} \bigcup JM_{u,v}$;
end
$J_{Global}$ = $JS_{Global} \bigcup JM_{Global}$;
return $J_{Global}$ ;
```

vector will be very long. Inspired by the theory that the subset of effective patterns should be adaptively learnt from the database [20], we employ a learning model containing three layers to obtain the optimal subset of CLBP-TOP patterns.

Layer 1 ensures the robustness of features with dominant pattern set, which is defined as the minimum set of pattern types covering δ (0<δ<1) of all patterns. Let p denote the total number of pattern types in the uth (u=1: XY, 2: XT, 3: YT plane) and $P_{u,\xi}$ the number of occurrences of pattern type ξ. The dominant pattern set of each orthogonal plane J_u can be derived with the following equation:

$$\begin{aligned} & J_u = arg \min |J_u| \\ & s.t. \frac{\sum_{\xi \in J_u} P_{u,\xi}}{\sum_{k=1}^{p} P_{u,k}} \geq \delta \end{aligned} \tag{2}$$

where $|J_u|$ denotes the number of elements in J_u. In this way, the most frequently occurring patterns in each plane are preserved which tend to be reliable to represent the structure of the plane. The rarely occurring patterns are removed for they probably come from interference and may result in a sparse histogram.

Layer 2 ensures the discriminative power of features. In order to minimize the within-class scatter, it is desired that examples belonging to the same class have same patterns. Therefore, intersection of dominant pattern sets is carried out across all training examples in the same class. Thus, the optimal subset of CLBP-TOP patterns learned from class c with n_c examples can be expressed as:

$$J_c = \{\bigcup_{u=1}^{3} \bigcap_{n=1}^{n_c} JS^n_u\} \bigcup \{\bigcup_{u=1}^{3} \bigcap_{n=1}^{n_c} JM^n_u\} \tag{3}$$

where JS^n_u and JM^n_u denote the dominant pattern set from uth plane of nth example with respect to sign and magnitude component, separately. The central pixel intensity component is not considered here for it makes less contribution than the other two components [13, 19].

Layer 3 constructs global dominant patten set. In the previous section, an example's image sequence has been divided into B blocks in spatial-temporal domain. To derive the global feature, J_c is extracted in each block B_v and then concatenate together as $J_{Global} = \bigcup_{v=1}^{B} J_{c,v}$. The learned J_{Global} from different classes then put together as the reference for feature extraction of testing sets.

With the learning model, CLBP-TOP is optimized by seeking out dominant pattern sets and minimizing the within-class scatter. Algorithm 2 shows the learning process of disCLBP-TOP with FSC and time division.

4. EXPERIMENTS AND ANALYSIS

4.1. Databases

Three databases SPOS [13], BBC [21], and UvA-NEMO [12] are specially collected for SVP recognition, details of which are demonstrated in Table I. Note that SPOS corpus contains natural color and infrared videos with only the onset phase of six basic expressions while the natural color videos of happy expression are employed here.

Table I: Details of BBC, SPOS and UvA-NEMO

Database	Subects	Total number	Resolution
	Female/Male	S/P	Frame rate
BBC	20	20	314× 286
	7/13	10 / 10	25fps
SPOS	7	80	640× 480
	3/4	66 / 14	25fps
UvA-	400	1240	1920× 1080
NEMO	185/215	597/ 643	50fps

The five facial key points are manually labeled to initialize tracking and facial regions are then aligned 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. LIBSVM [22] is employed as the classifier with one class of spontaneous and another class of posed. Besides, linear kernel is adopted in consideration of its good performance, simplicity and, last but not least, the speed.

4.2. Experiment A: Comparison of Different Space and Time Division Methods

As the contribution evaluation of eye, cheek and mouth regions for SVP smile recognition has been done in [12], here, we focus on the effect of different space and time division. 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 Rise-Sustain-Decay (R-S-D), θ is assigned to 0.9. The experiments are implemented on UvA-NEMO here. In Table II, $H \times H$ 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 of 85.26% is achieved with H = 8 and T = 3, which shows over-dividing and coarse dividing are undesirable. Over-dividing makes the statistics of local texture invalid while coarse dividing can not well construct global structure. T is equal to 3 when combined with FSC and H is equal to 8 combined with R-S-D. FSC performs better than $H \times H$, verifying the advantages of FSC mentioned. 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 T EQUAL combined with FSC while the effect is not very obvious with $H \times H$. As timing information for SVP smile are different [16] and different facial subregions are at different states at each moment, R-S-D properly divides the timing of cropped facial subregions according to intensity while T EQUAL may wrongly cuts the information.

Table II: Comparison of different space and time division methods using disCLBP-TOP on UvA-NEMO

Facial region cropping	Time division	Accuracy(%)
FSC	R−S−D	91.40
FSC	T EQUAL	87.54
$H \times H$	R−S−D	85.03
$H \times H$	T EQUAL	85.26

4.3. Experiment B: Effect of Rise, Sustain and Decay

To explore the effect of different phases of a smile, the feature of each phase are used separately. SVM is utilized to classify the feature of each phase individually and the voting rule is employed to fuse results of classifiers. As shown in Fig. 3, the rise and the sustain phase achieves higher accuracy than the decay phase which is close to the result obtained with onset-apex-offset division in [12]. However, the sustain phase performs better than the rise in our experiment while the recognition rate of apex is lower than onset in [12] . The reason is that the sustain phase includes some frames from onset and offset phases using our division. The combination of rise and sustain phase achieves close accuracy compared to all phases employed, which shows the decay phase plays less important role in SVP smile recognition. Back to the extracted features, it is found that the number of discriminative pattern types extracted in the decay phase are less than the other two phases, which is the essential reason of above phenomenon.

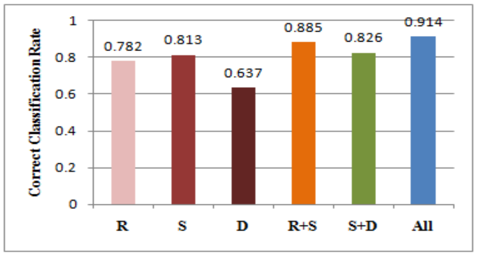

Fig. 3: Evaluation of different smile phases on UvA-NEMO

4.4. Experiment C: Comparison with Other Methods

The comparison of our method with the state-of-the-art SVP smile recognition methods proposed in [12, 13] is made on three databases with the same experimental protocols and disCLBP-TOP and CLBP-TOP are implemented with FSC and R-S-D division. Correct recognition rates are given in Table III, which show our proposed approach outperforms others. DisCLBP-TOP achieves better results than CLBP-TOP which validates its discriminative and robust power. The recognition rate on SPOS is lower than other two is because SPOS only contains the onset phase of smiles

Table III: Correct recognition rates on the three databases

Methods	Correct Classification Rate (%)		
	BBC	UvA-NEMO	SPOS
disCLBP-TOP	**90.00**	**91.40**	**79.50**
CLBP-TOP	80.00	83.03	71.50
Dibeklioğlu et al. [12]	90.00	87.02	75.00
Pfister et al. [13]	70.00	73.06	67.50

5. CONCLUSIONS

In this paper, we propose an appearance-based feature extraction method for SVP smile recognition instead of geometry-based ones. Experimental results show that the recognition rate is improved using the proposed discriminative local spatial-temporal descriptor: disCLBP-TOP, which confirms its robustness and discriminability. Besides, flexible facial subregion cropping (FSC) shows better performance than equal division in the spatial domain since it takes into account specific subregions, facial organ size of people and reduction of redundant information. In temporal domain, our proposed time division method achieves the best result when combined with FSC, which shows the advantage of time division according to smile dynamics. Experiments on three benchmark databases also show our proposed approach outperform the state-of-the-art methods. Our work could pave the way for computers that better assess the emotional states of their users. It could also be used as a tool for the analysis of smiles in psychology or help people with difficulty in interpreting expression.

6. REFERENCES

[1] Mark G Frank and Paul Ekman, "Not all smiles are created equal: The differences between enjoyment and nonenjoyment smiles," *International Journal of Humor Research*, vol. 6, no. 1, pp. 9–26, 1993.

[2] Simon Baron-Cohen, Howard A Ring, Edward T Bullmore, Sally Wheelwright, Chris Ashwin, and SCR Williams, "The amygdala theory of autism," *Neuroscience & Biobehavioral Reviews*, vol. 24, no. 3, pp. 355–364, 2000.

[3] Beat Fasel and Juergen Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[4] Maja Pantic and Leon J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1424–1445, 2000.

[5] Jeffrey F Cohn and Karen L Schmidt, "The timing of facial motion in posed and spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 2, pp. 121–132, 2004.

[6] Karen L Schmidt, Sharika Bhattacharya, and Rachel Denlinger, "Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises," *Journal of Nonverbal Behavior*, vol. 33, no. 1, pp. 35–45, 2009.

[7] Karen L Schmidt, Zara Ambadar, Jeffrey F Cohn, and L Ian Reed, "Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling," *Journal of Nonverbal Behavior*, vol. 30, no. 1, pp. 37–52, 2006.

[8] M Hoque and Rosalind W Picard, "Acted vs. natural frustration and delight: Many people smile in natural frustration," in *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG)*, pp. 354–359. 2011.

[9] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[10] Michel F Valstar, Hatice Gunes, and Maja Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *International Conference on Multimodal Interfaces*, pp. 38–45. 2007.

[11] M Hoque, Daniel McDuff, and R Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 323–334, 2012.

[12] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers, "Are you really smiling at me? spontaneous versus posed enjoyment smiles," in *International Conference on Computer Vision (ECCV)*, pp. 525–538. 2012.

[13] Tomas Pfister, Xiaobai Li, Guoying Zhao, and M Pietikainen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 868–875. 2011.

[14] Hong Liu and Pingping Wu, "Comparison of methods for smile deceit detection by training au6 and au12 simultaneously," in *IEEE International Conference on Image Processing (ICIP)*, pp. 1805–1808. 2012.

[15] Hamdi Dibeklioglu, Roberto Valenti, Albert Ali Salah, and Theo Gevers, "Eyes do not lie: spontaneous versus posed smiles," in *International Conference on Multimedia*, pp. 703–706. 2010.

[16] Zara Ambadar, Jeffrey F Cohn, and Lawrence Ian Reed, "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous," *Journal of Nonverbal Behavior*, vol. 33, no. 1, pp. 17–34, 2009.

[17] Michel F Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F Cohn, "Spontaneous vs. posed facial behavior: automatic analysis of brow actions," in *Proceedings of the 8th international conference on Multimodal interfaces*, pp. 162–170. 2006.

[18] Guoying Zhao and Matti Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 915–928, 2007.

[19] Zhenhua Guo, Lei Zhang, and David Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1657–1663, 2010.

[20] Yimo Guo, Guoying Zhao, and Matti Pietikäinen, "Discriminative features for texture description," *Pattern Recognition*, vol. 45, no. 10, pp. 3834–3843, 2012.

[21] "BBC-dataset," Available at `http://www.bbc.co.uk/science/humanbody/mind/surveys/smiles/`.

[22] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.