Spatio-Temporal Texture-Based Feature Extraction for Spontaneous Facial Expression Recognition

Siti Khairuni Amalina Kamarol, Nor Syazana Meli, Mohamed Hisham Jaward, Nader Kamrani
Electrical and Computer Systems Engineering, School of Engineering
Monash University Malaysia
46150 Selangor, Malaysia
siti.khairuni@monash.edu, norsyazanameli@gmail.com,
mohamed.hisham@monash.edu, nader.kamrani@monash.edu

Abstract

Recently, recognition of naturalistic expressions known as spontaneous facial expressions has attracted attention from researchers due to its significant application in behavioral and clinical research. Currently, most of the work consider recognition of posed expressions. In this paper, we propose a spatio-temporal feature extraction method, Spatio-Temporal Texture Map (STTM), for recognition of spontaneous expressions and compare its performance against that of state-of-the-art feature extraction methods. Both appearance-based and geometry-based feature extraction approaches are considered for comparisons against STTM. The appearance-based techniques considered are Volume Local Binary Pattern (VLBP) and Local Binary Pattern from Three Orthogonal Planes (LBP-TOP) whereas a multi-view tree-based face detector is considered as a geometry-based technique. Support Vector Machine (SVM) is used as the classifier where the extracted features are classified into classes of naturalistic expressions. The feature extraction methods are evaluated over the spontaneous facial expression data from CASME II database. Experimental results show that STTM is capable of recognizing spontaneous expressions and outperforming the other methods in terms of recognition rate, accuracy and computational cost.

1 Introduction

According to A. Mehrabian [1], information from verbal communication is contributed by three channels which are spoken language (7%), voice intonation (38%) and facial expression of the speaker (55%). This shows the importance of facial expression analysis and explains why facial expression recognition (FER) is a growing area of research since the past two decades [2]. [1] also listed a few factors contributing to change in facial expressions: mental states, verbal and nonverbal communication, and physiology. Each of these factors triggers different intensities of facial expression making the task of recognizing a spontaneous expression more challenging. With the widespread usage of computers in society, a more natural and friendly interaction between computer and human beings is needed as it could be useful in many applications such as surveillance, criminal identification, and psychological studies [2, 3].

Within the past two decades, most existing FER research have been focusing on deliberate expressions, known as posed facial expressions [4]. However, in real life situations, most of the time people tend to express their emotion spontaneously. Therefore, for practical purposes, FER systems should be developed to recognize natural expressions as well. FER systems are built based on three main components, namely face detection or tracking, facial feature extraction, and facial expression classification [2]. In general, feature extraction algorithms can be categorized into two: appearance-based and geometry-based. Appearance features are based on texture variations in images. Texture analyzing field is dominated by Local Binary Patterns (LBP) based operators as it is proven that these operators perform better than most existing methods [5]. Volume Local Binary Patterns (VLBP) is one of the appearance-based operators introduced by Zhao and Pietikäinen [6] in 2007. It is an extension to the basic LBP which enables extraction of features from video sequences by taking into account both time and spatial domains. VLBP is then extended to Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [6] which reduces the computational complexity of VLBP. Geometry-based features consist of landmark points annotated on the face. A multi-view tree-based face detector capable of detecting landmark points was proposed in [7]. It is claimed to be rotation invariant and able to detect faces from several viewpoints. Given an image, this algorithm automatically annotates 68 landmark points on the face.

In this paper we propose an appearance-based feature extraction algorithm, Spatio-Temporal Texture Map (STTM), for video sequences. STTM extracts spatio-temporal appearance information using an extension of the standard Harris corner function [8]. Laptev [9] has applied three-dimensional Harris corner function to detect spatio-temporal interest points for recognition of human activities. In STTM, instead of detecting interest points, the three-dimensional Harris corner function is used to extract spatio-temporal texture map from video sequences. It enables subtle motion patterns on the face to be captured with low computational cost. Motivated by the success of STTM on posed expressions [10], we propose an approach based on STTM and compare it with the rest of the aforementioned appearance-based and geometry-based feature extraction methods on video sequences depicting spontaneous expressions. One of the major contributions of this work is to compare the performance of state-of-the-art appearance-based feature extraction algorithms, namely VLBP, LBP-TOP, and STTM, for spontaneous video sequences. Our second main contribution is to identify the possible best feature extraction algorithms for recognizing spontaneous facial expressions. First, we evaluate the per-

formance of appearance-based methods: VLBP, LBP-TOP, and STTM. Then, we study the performance of the geometry-based approach by comparing features extracted only from apex frame of each video with features extracted from apex and neighboring frames of the video.

This paper is organized as follows: In Section 2, we describe the proposed feature extraction algorithm, STTM. In Section 3, we describe the experimental procedures, present the experimental results and discuss the performance of the algorithms for recognizing spontaneous expressions. Finally, Section 4 concludes the paper and proposes future work.

2 Proposed Facial Feature Extraction

The proposed feature extraction method, STTM, detects regions in a video sequence f where the pixel intensities vary significantly in both space and time domains. First, we model f by its linear scale-space representation L constructed by convoluting f with a Gaussian kernel.

$$L(\cdot;\sigma_l^2,\tau_l^2) = g(\cdot;\sigma_l^2,\tau_l^2) * f(\cdot) \tag{1}$$

where σ_l^2 and τ_l^2 are the spatial variance and temporal variance of the spatio-temporal Gaussian kernel g which is defined as

$$g(x,y,t;\sigma_l^2,\tau_l^2) = \frac{\exp(-(x^2+y^2)/2\sigma_l^2 - t^2/2\tau_l^2)}{\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}} \tag{2}$$

Förstner [11] and Harris [8] suggested that given an image, distinct points can be detected by considering a Gaussian window in the image. Shifting the window by a small amount in various directions provides locations in the image where pixel intensities vary significantly in space domain. Since we are considering the whole video sequence as input, our method finds locations with significant pixel intensity variations in both space and time domains. Such locations are determined by convolution of a Gaussian weighting function $g(\cdot;\sigma_i^2,\tau_i^2)$ with a spatio-temporal second-moment matrix as follows:

$$\mu = g(\cdot;\sigma_i^2,\tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \tag{3}$$

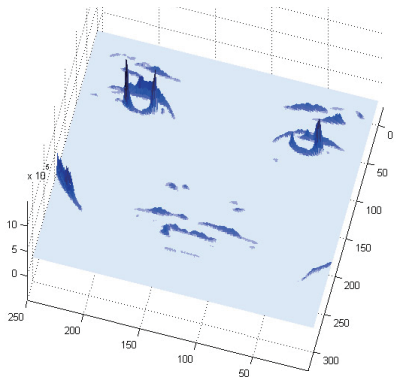

Figure 1: A spatio-temporal texture map of the last frame of a video sequence depicting surprise.

The second-moment matrix consists of first order spatial and temporal derivatives of L defined as

$$L_x(\cdot;\sigma_l^2,\tau_l^2) = \partial_x(g * f),$$
$$L_y(\cdot;\sigma_l^2,\tau_l^2) = \partial_y(g * f),$$
$$L_t(\cdot;\sigma_l^2,\tau_l^2) = \partial_t(g * f),$$

where $\sigma_i^2 = s\sigma_l^2$, $\tau_i^2 = s\tau_l^2$, and s is a constant.

A spatio-temporal texture map of f is then obtained using the three-dimensional Harris corner function constructed by combining the determinant and trace of μ as follows:

$$\begin{aligned} H &= \det(\mu) - k\ \mathrm{trace}^3(\mu) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3 \end{aligned} \tag{4}$$

where λ_1, λ_2, and λ_3 are the significantly large eigenvalues of μ and k is a constant. We use H as a texture map which shows pixel intensity variations of f in space and time domains. The texture map is then normalized to remove the effect of illumination variations of the images. Figure 1 shows a texture map of the last frame of a video sequence depicting surprise. A block-based feature representation [6] is then implemented where features from a video sequence are extracted from smaller blocks and a histogram is constructed for each block. Histograms from the entire video sequence are then concatenated, forming a feature vector representing the video. In order to better capture the subtle variations in f, we implement A-law compression [12] before computing a histogram.

3 Experiments and Results

In this section, we describe the experimental procedures and discuss the experimental results. Before extracting the appearance-based features using VLBP, LBP-TOP, and STTM, video frames are cropped using Viola-Jones face detector [13]. This step extracts the desired facial region and removes the background for better recognition. A block-based feature representation [6] is then implemented to obtain feature vectors. Since the multi-view tree-based detector used to extract geometric features is able to detect faces automatically, uncropped video frames are used. The features are then classified using one-against-one SVM [14]. All experiments were carried out using 4-fold cross validation. The optimal parameters obtained are $L = 3$, $P = 2$, $R = 3$ for VLBP, $P_{XY} = P_{XT} = P_{YT} = 8$, $R_X = R_Y = R_T = 3$ for LBP-TOP, and $\sigma_l^2 = 2$,

Table 1: Number of video sequences in CASME II dataset.

Facial Expression	No. of Videos	No. of Frames
Disgust	57	3822
Happiness	30	2208
Repression	26	2066
Sadness	6	234
Surprise	25	1605
Total	144	9395

Table 2: Recognition rates for appearance-based features.

Facial Expression	VLBP	LBP-TOP	STTM
Disgust	100.00	100.00	100.00
Happiness	100.00	93.33	100.00
Repression	96.15	96.15	96.15
Sadness	66.67	83.33	100.00
Surprise	84.00	84.00	96.00
Average	89.36	91.36	98.43

Table 3: Number of video sequences for subjects without eyeglasses in CASME II dataset.

Facial Expression	No. of Videos	No. of Frames
Disgust	32	2125
Happiness	18	1262
Repression	21	1728
Surprise	14	863
Total	85	5978

$\tau_l^2 = 2$, $k = 0.04$ for STTM. Varying the number of blocks gives us the optimal number of blocks for feature extraction which is 9×9.

3.1 Evaluation on appearance-based features

We performed evaluation of VLBP, LBP-TOP and STTM on video sequences depicting five different expressions from CASME II dataset [15]. These spontaneous expressions are: disgust, happiness, repression, sadness, and surprise. The number of video sequences and frames for each expression are tabulated in Table 1.

The recognition rates for all three methods are displayed in Table 2. It can be observed that the lowest recognition rates achieved by VLBP and LBP-TOP are for the case of sadness with 66.67% and 83.33% respectively. STTM, on the other hand, achieved the highest recognition rate for every expression and the highest average recognition rate of 98.43%. Despite the small number of videos available for training and testing for the case of sadness, STTM has shown a superior performance with 100% recognition rate while the recognition rates for VLBP and LBP-TOP for that expression are the lowest among all expressions.

Based on the results, it is noted that the amount of training data plays an important role in achieving high recognition results. In general, the recognition rates obtained by all three approaches are considered high and they are indeed efficient to be used in analyzing both posed and spontaneous facial expressions.

3.2 Evaluation on geometry-based features

In the second part of the experiments, we analyzed the performance of geometry-based technique on video sequences depicting four different expressions, consisting of disgust, happiness, repression, and surprise. The number of video sequences and frames in each expression are tabulated in Table 3. Note that the number of video sequences used in this part of the experiments is much lower. This is due to the weakness of the geometry-based method to detect landmark points accurately for subjects wearing eyeglasses. Therefore for fair comparisons with the other techniques, in this experiment we only considered subjects without eyeglasses.

Two cases are considered in this experiment. The first case is extracting landmark points using multi-view tree-based algorithm from the apex frame of each video. The features are static where there is no variation in the time domain. The second case is extracting landmark points from the apex frame and its neighboring frames (one frame before and after apex frame), hence capturing dynamic information. Results obtained are tabulated in Table 4. The results demonstrate that considering the neighboring frames in addition to the apex frame increases recognition rates. This is due to the fact that the features contain more useful information for better recognition. Overall, considering neighboring frames has slightly improved the recognition performance.

Table 4: Recognition rates for geometry-based features for subjects without glasses.

Facial Expression	Apex Frame	Apex and Neighboring Frames
Disgust	96.88	96.88
Happiness	83.33	72.22
Repression	85.71	90.48
Surprise	73.33	86.67
Average	84.81	86.56

Table 5: Recognition rates for appearance-based features for subjects without eyeglasses.

Facial Expression	VLBP	LBP-TOP	STTM
Disgust	93.75	96.88	90.63
Happiness	72.22	83.33	83.33
Repression	95.24	100.00	100.00
Surprise	85.71	85.71	92.86
Average	86.73	91.48	91.71

3.3 Evaluation on appearance-based and geometry-based features

The third part of the experiments aims to evaluate the performance of all the appearance-based and geometry-based features for spontaneous expression recognition. For fair comparisons, only video sequences of subjects without eyeglasses are considered in this experiment. Details of the data used are shown in Table 3. Since it has been proved that considering neighboring frames in addition to the apex frame of each video produces slightly better recognition, we have chosen this approach to represent geometry-based features. The results obtained are shown below in Table 5. Besides evaluating performance in terms of recognition rates, we also evaluate the computational cost of these

Table 6: Comparisons of computational cost.

No. of Frames	Computational Time (s) Appearance-based			Geometry-based
	VLBP	LBP-TOP	STTM	Multi-view tree-based
5978	15405	16165	9164	30858
1	2.577	2.704	1.533	5.162

algorithms. The time taken by these algorithms to process the data are shown in Table 6. Experiments were performed on an Intel Xeon 3.5GHz workstation with 16GB RAM.

Based on the results in Table 4 and Table 5, we can observe that LBP-TOP and STTM achieved high recognition rates for most of the expressions with the average of 91.48% and 91.71% respectively. The difference in recognition rates for the algorithms are quite small. The recognition rates for all the appearance-based algorithms are lower for most of the expressions compared to the case where more videos were used (Table 2). This is due to insufficient number of frames used in training. As mentioned before, based on initial testing the multi-view tree-based algorithm performs better on data with no occlusion on the face region such as eyeglasses. One of the limitations of current research in spontaneous FER is lack of data available for analysis. Due to the difficulty of designing an environment for spontaneous data collection, there are few spontaneous facial expression datasets publicly available.

In terms of computational cost, STTM outperforms the other methods with the shortest time taken to process the data. On average, STTM took only 1.533 seconds to perform computations for one frame while VLBP and LBP-TOP took more than 2 seconds. The geometric features required the longest time to perform computations.

In general, based on the performance of the algorithms, we can see that all methods perform relatively well on spontaneous dataset despite the small variations produced by the expressions. However, based on the performance of STTM on the entire dataset and based on the computational cost of the algorithm we can say that STTM is the most efficient feature extraction algorithm compared to the rest. Its fast computation makes it feasible for real-time applications.

4 Conclusion

In this paper, we proposed a spatio-temporal feature extraction method for facial expression recognition and performed comparative analysis against state-of-the-art algorithms. We performed experiments using video sequences from CASME II, a spontaneous dataset. Based on the results and discussion, we conclude that given enough data for analysis, STTM is the most efficient feature extraction algorithm outperforming the other methods in terms of recognition rate and computational time. Hence it is reliable for achieving high accuracy and fast processing time. As future work, we plan to implement these algorithms on a larger spontaneous dataset consisting of subjects from different ethnicity, skin color, age and other external factors.

Acknowledgment

This work was supported by Malaysia Ministry of Higher Education Fundamental Research Grant Scheme, FRGS/1/2012/TK02/MUSM/03/2.

References

[1] A. Mehrabian, "Communication with words," *Psychology Today*, vol. 2, no. 4, pp. 53-56, 1968.

[2] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259-275, 2003.

[3] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 3, pp. 311-324, 2007.

[4] S. He, S. Wang, and Y. Lv., "Spontaneous facial expression recognition based on feature point tracking," in *6th Int. Conf. Image and Graphics*, Hefei, Anhui, 2011, pp. 760-765.

[5] J. Bihan, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *IEEE Int. Conf. Automatic Face & Gesture Recognition and Workshops*, Santa Barbara, CA, 2011, pp. 314-321.

[6] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915-928, 2007.

[7] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, 2012, pp. 2879-2886.

[8] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vison Conf.*, Manchester, 1998, pp. 147-152.

[9] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. 9th IEEE Int. Conf. Computer Vision*, Nice, France, 2003, pp. 147-152.

[10] S. K. A. Kamarol, M. H. Jaward, J. Parkkinen, R. Parthiban "Spatio-temporal feature extraction for facial expression recognition," submitted to IET Image Processing.

[11] W. Förstner and E. Gülch, "A fast operator for detection and precise location of distinct points, corners and centres of circular features," in *Proc. ISPRS Intercommission Conf. Fast Processing of Photogrammetric Data*, Interlaken, 1987, pp. 281-305.

[12] A. Gersho, "Principles of quantization," *IEEE Trans. Circuits and Systems*, vol. 25, no. 7, pp. 427-436, 1978.

[13] P. Viola and M. Jones, "Robust real-time face detection," *Int. Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, 2004.

[14] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415-425, 2002.

[15] W. J. Yan, X. Li, S. J. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, 2014.