AUTOMATIC FACIAL EXPRESSION RECOGNITION FOR AFFECTIVE COMPUTING BASED ON BAG OF DISTANCES

Fu-Song Hsu[1]*,Wei-Yang Lin*[2]*,Tzu-Wei Tsai*[3]

[1,2]Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan, R.O.C.
E-mail: {hfs95p,wylin}cs.ccu.edu.tw
[3]Department of Multimedia Design, National Taichung University of Science and Technology, Taichung, Taiwan, R.O.C.
E-mail: wei@nutc.edu.tw

ABSTRACT

In the recent years, the video-based approach is a popular choice for modeling and classifying facial expressions. However, this kind of methods require to segment different facial expressions prior to recognition, which might be a challenging task given real world videos. Thus, in this paper, we propose a novel facial expression recognition method based on extracting discriminative features from a still image. Our method first combines holistic and local distance-based features so that facial expressions could be characterized in more detail. The combined distance-based features are subsequently quantized to form mid-level features using the bag of words approach. The synergistic effect of these steps leads to much improved class separability and thus we can use a typical method, e.g., Support Vector Machine (SVM), to perform classification. We have performed the experiment on the Extended Cohn-Kanade (CK+) dataset. The experiment results show that the proposed scheme is efficient and accurate in facial expression recognition.

Index Terms— Facial expression recognition, facial features, bag of words, Affective Computing

I. INTRODUCTION

In this rapid development of advanced computer technology era, automatic facial expression recognition has gained an increasing interest in building intelligent environments. The facial expression recognition method can be adopted as the central mechanism in sensing the emotions of humans using a camera, linking the camera space to media content. To simulate virtual objects to real activities, the system must have the ability to immediately handle and respond to the captured physiological signals. Hence, creating an effective facial expression recognition method is the primary goal of this study.

The facial expression recognition methods can be divided into two main categories based on data sources: sequence and image [1]. In sequence-based methods, one expression is characterized with an image sequence. The image sequence includes the onset to peak formation of the facial expressions. The image-based methods use only the current image to recognize the expression of the image. Therefore, facial expression recognition using image-based methods is more difficult than using sequence-based because less information is available.

Although the sequence-based input provides rich facial motions, the recognition of facial expressions has remained a challenging and sophisticated task due to several reasons. Firstly, there is a need to select a neutral face first to serve as a reference image [2]. If the reference image is not correctly selected, the recognition task cannot be expected to perform well. Secondly, expressions often last for various length of time. However, the system performance depends on the appropriate segmentation from a live streaming [3]. Sophisticated segmentation methods can be processed at a much higher computational power, but it's better to avoid them in real environments.

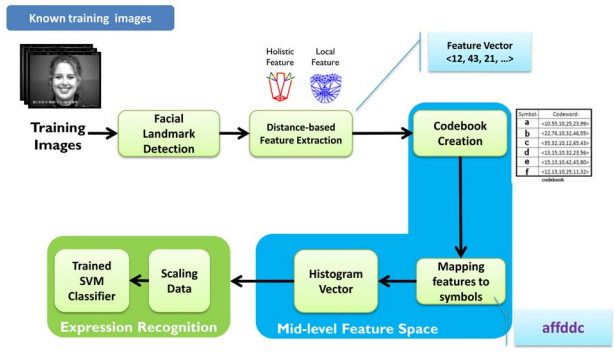
To resolve the above mentioned issues, we devote our attention to use a still image as input. The image-based methods use only the current image to recognize the expression of the image. Therefore, we do not need to select the onset to peak formation of the facial expressions. The key issue is how to extract expressive and discriminating facial features from a still image. This study focuses on the use of distance-based features to deal with spatial variations associated with human expressions. In order to have rich facial features, we propose to combine the holistic feature [4] and local feature [5] so that a facial shape can be characterized in more detail. These two features provide complementary information and are integrated to yield a better discriminative power. Then, we apply the k-means algorithm to create a codebook that learning from large volume of training samples. Thus, the distance-based features can be quantized into codewords to represent facial features in the mid-level feature space. Once expressions have been converted into strings, the problem of expressions recognition can be quantified by calculating distance between strings. Here, the proposed system utilizes histogram-based vectors to train a Support Vector Machine (SVM) classifier is a simple, yet efficient solution for recognizing human expressions. Experiment results demonstrate

that our approach can accurately and efficiently recognize facial expressions.

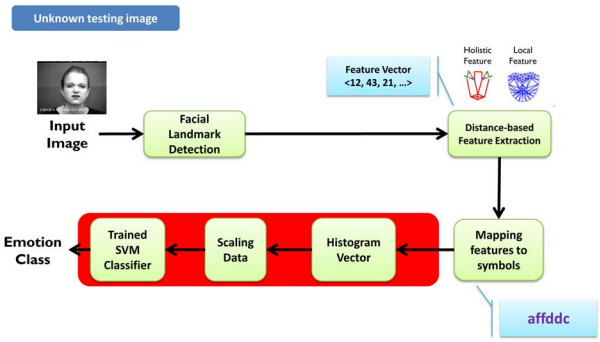
II. PROPOSED METHOD

II-A. Overview of the Proposed Method

The proposed system basically consists of four parts, including facial landmark detection, distance-based feature extraction, mid-level feature space and expression recognition. The flowcharts of training and recognition processes in our system are shown in Figure 1.



(a) Process flow of training;



(b) Process flow of recognition.

Fig. 1. Flowchart of the proposed system.

II-B. Facial Landmark Detection

For facial landmark detection, we employ an Active Appearance Models (AAM) [6] to locate landmarks in a face image. The landmarks have been tracked with an AAM fitting using a default model. There are 68 facial points in the default model for both the x and y coordinates. Once we have tracked a face by using the AAM fitting, we can use these facial landmark to derive its features.

II-C. Distance-based Feature Extraction

Given an input image, the AAM can be employed to turn the high dimensional pixel data into a lower dimensional landmarks vector as $\mathbf{S} = [x_1; y_1; x_2; y_2; ...; x_p; y_p]$. Here, we explain how to extract the distance-based features from these facial points.

1) Local Features

The local distances is computed from a set of landmarks $\mathbf{S}$, in $\mathbf{A} \in \mathbb{R}^2$, is a triangle network whose vertices are the landmarks and follow the Delaunay triangulation [7]. The Delaunay triangulation is the unique triangulation such that no vertex is inside its circumcircle. Given a set $\mathbf{P}$ of points in the Euclidean space, the graph of the Delaunay triangulation can be defined by the

$$DT(\mathbf{P}) = \det \begin{pmatrix} 1 & A_x & A_y & A_x^2 + A_y^2 \\ 1 & B_x & B_y & B_x^2 + B_y^2 \\ 1 & C_x & C_y & C_x^2 + C_y^2 \\ 1 & D_x & D_y & D_x^2 + D_y^2 \end{pmatrix} > 0 \quad (1)$$

where $DT(\mathbf{P})$ is to detect if point D lies inside the circumcircle of A, B, C; this determinant is positive. Figure 2(a) shows the Delaunay triangulation result on the landmarks $\mathbf{S}$. Next, let $\{d_i^{DS}\}_{i=1}^{\varepsilon}$ be the local distances that contains all edges ε of the Delaunay triangulation. The distance is normalized by

$$\overline{d}_i^{DS} = \frac{d_i^{DS}}{\sum_{j=1}^{\varepsilon} d_j^{DS}} \quad (2)$$

Previous research have addressed [4], [8], [9] the eyebrows, eyes, and mouth points have a strong relation to the information about the shown facial expression and hence we adopt local distances mostly came from these regions. The length of the local features is standardized as follows

$$\hat{d}_i^{DS} = \overline{d}_{i \times \zeta}^{DS}, \forall i \in [1 \ldots \zeta] \quad (3)$$

where ζ denotes the number of selected distances on a face.

2) Holistic Features

Humans usually recognize emotions according to both global facial appearance and local variations of facial regions. In this work, we attempt to consider holistic and local representations of a face simultaneously. In general, holistic representations uses the whole face and focus on the facial variations of crossing local regions. Tanchotsrinon *et al.* [4] demonstrate that the graph-based features of the whole face carry valuable information for recognizing emotions. We agree with Tanchotsrinon that such pattern feature can help to recognize emotions efficiently and hence we adopt it as holistic representation in our system.

As shown in Figure 2(b), the distance information between adjacent regions have been examined, such as calculated from the eyebrow to eye. Since different people have different face size, the holistic features must also be normalized to reduce the effect of scale

(a) The illustration of local distances

(b) The illustration of holistic distances

Fig. 2. Distance-based feature extraction

variations. The normalization of the holistic features is similar to that of the local features. After that, we simply join the two features together, called the combined feature.

II-D. Mid-level feature space

In the mid-level feature space, the key components include (1) codebook creation, (2) mapping features to symbols, and (3) histogram vectors. The details are described as follows.

1) Codebook Creation: Given an input image, expression recognition is performed by ranking the facial features in the training set according to their similarity to the input image. To simplify the analysis, the extracted features are then quantized using a codeword. The codeword is commonly learned by clustering k random means $m_1, \ldots, m_k$ of feature descriptors. Assign each distance d to the cluster with the closest mean:

$$S_i^t = d_p : \|d_p - m_i^t\| \leq \|d_p - m_j^t\|, \forall 1 \leq j \leq k \quad (4)$$

where each d_p goes into exactly one S_i^t, so that we can calculate the new means in the $t+1$ interaction by (5)

$$m_i^{t+1} = \frac{1}{N_i} \sum_{d_i \in S_i^t} d_i \quad (5)$$

where N_i denotes the number of samples in S_i^t. The algorithm is deemed to have converged when the m_i^{t+1} equals m_i^t.

2) Mapping features to symbols: Once a codebook is built, we can assign each subset of extracted features Z to one of several k codewords as a symbol. The classification rule is based on a quantity in L2-norm, defined as

$$d_{\hat{k}} = \min_{1 \leq k} \bar{d}_k(Z), \quad (6)$$

with

$$\bar{d}_k(Z) = \frac{\sum_{i=1}^{n} \|Z_i - d_i\|}{n} \quad (7)$$

where n is the number of samples in the subset. Thus, a combined feature is represented by codewords.

3) Histogram vectors: To calculate the distance of two strings, we define h be a histogram of k bins associated with a representative feature. Then the h is used to encode the distribution of a string. The normalized histogram $\bar{h}$ is given by (8) and therefore the sum of all bins are equal to 1.

$$\bar{h}(i) = \frac{h(i)}{\sum_{i=1}^{k} h(i)}, \quad (8)$$

II-E. Expression Recognition

In this study, we exploit the BoW approach for representing facial expressions. Each facial expression looks like a bag, which contains some words from the codebook. Once facial expressions have been converted into codewords, *i.e.* string, the category of an input facial expression can be determined by string distance measuring.

In this setting, the nearest neighbor rule seems to be a natural choice, classifying the input data to the category of its nearest neighbor. However, nearest neighbor searching is very time-consuming. Therefore, we construct a classifier to learn the distributions of each category. In this work, we use the SVM classifier instead of using nearest neighbor rule. Our SVM classifier is trained using the libSVM [10].

III. EXPERIMENTAL RESULTS

In this study, we conduct a series of experiments to evaluate the proposed method. Experiments are conducted on the Extended Cohn-Kanade (CK+) dataset, which is the largest dataset commonly used on this problem. The CK+ dataset contains 327 emotion labeled image sequences, which are AAM tracked with 68 points landmarks for each image. There are seven facial expression categories, namely happy, sadness, anger, fear, disgust, contempt, and surprise. Here, we only use the last frame from an image sequence to perform expression recognition.

III-A. Comparison of different features

We first investigate the benefits of combing different features in expression classification. Hence, we have conducted experiments to evaluate the discrimination capability of three different features, namely local feature, global feature, and the combined feature. Table I compares different types of distance-based features. Results show that the combined feature achieves the best performance with a rating of 87.7%, outperforming the other features. Figure 3 shows the confusion matrices of the local feature, holistic feature, and combined feature. It can be observed from Figure 3 that the combined feature is usually more discriminative than individual features for each expression category.

The proposed method is implemented using Visual Studio C++ 2008 and evaluated on an Intel Core2 CPU-P8600 2.4 GHz Laptop with 4 GB memory. The classification procedure works at about 35 frames per second on the CK+ dataset. Notice that our system requires detailed facial

landmarks to be extracted, which might cause an increase in the time of evaluation of real-world videos.

Table I. Expression recognition using different features (using 128 codewords)

Distance-based Features	Recognition rate (%)
local feature	79.2
holistic feature	77.0
combined feature	87.7

	Happy	Sadness	Anger	Fear	Disgust	Contempt	Surprise	Total
Happy	62	0	1	4	1	1	0	69
Sadness	1	12	8	1	4	2	0	28
Anger	1	4	33	0	5	2	0	45
Fear	6	1	0	13	1	1	3	25
Disgust	0	2	2	3	52	0	0	59
Contempt	2	3	3	1	1	8	0	18
Surprise	1	1	0	1	0	1	79	83
Total								327

(a)

	Happy	Sadness	Anger	Fear	Disgust	Contempt	Surprise	Total
Happy	64	0	1	1	1	2	0	69
Sadness	0	8	9	2	6	3	0	28
Anger	2	3	27	0	12	1	0	45
Fear	3	3	0	18	0	0	1	25
Disgust	1	4	4	0	48	2	0	59
Contempt	0	4	0	2	4	7	1	18
Surprise	0	0	0	0	2	1	80	83
Total								327

(b)

	Happy	Sadness	Anger	Fear	Disgust	Contempt	Surprise	Total
Happy	66	0	1	1	1	0	0	69
Sadness	0	21	3	1	3	0	0	28
Anger	0	4	37	0	3	1	0	45
Fear	1	2	0	18	1	1	2	25
Disgust	0	1	2	1	54	0	1	59
Contempt	0	3	3	1	0	11	0	18
Surprise	0	1	0	0	1	1	80	83
Total								327

(c)

Fig. 3. (a) Confusion matrix of local feature. (b) Confusion matrix of holistic feature. (c) Confusion matrix of combined feature.

IV. CONCLUSIONS

Automatically recognizing facial expressions is important to understand human emotion and to related applications, such as multimodal user interfaces. A novel framework for recognizing facial expressions is presented in this paper. We summarize the advantages of the proposed system as follows.

- We introduce novel distance-based features to provide complementary information and are integrated to yield a better discriminative power.
- The BoW model is applied to learn the training sequences and to construct a codebook automatically.
- Experiment results that our approach can accurately and efficiently recognize facial expressions.
- The system uses only the current image to recognize the expression of the image, making it more suitable for real-time

In our future work, we will improve the accuracy of the recognition rate, and give more comparison results on different datasets to further verify the performance of our method.

V. REFERENCES

[1] F. De la Torre and J. F. Cohn, *Guide to Visual Analysis of Humans: Looking at People*, chapter Facial Expression Analysis, Springer, 2011.

[2] C.-C. Lee, S.-S. Huang, and C.-Y. Shih, "Facial affect recognition using regularized discriminant analysis-based algorithms," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, 2010.

[3] A. Licsar, T. Sziranyi, L. Kovacs, and B. Pataki, "A folk song retrieval system with a gesture-based interface," *MultiMedia, IEEE*, vol. 16, no. 3, pp. 48 –59, july-sept. 2009.

[4] C. Tanchotsrinon, S. Phimoltares, and S. Maneeroj, "Facial expression recognition using graph-based features and artificial neural networks," in *Proc. IEEE Conf. on Imaging Systems and Techniques (IST)*, 2011, pp. 331–334.

[5] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 94–101.

[6] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681 –685, jun 2001.

[7] B. Delaunay, "Sur la sphère vide," *Bulletin of Academy of Sciences of the USSR*, , no. 6, pp. 793–800, 1934.

[8] M.-C. Su, Y.-J. Hsieh, and D.-Y. Huang, "Facial expression recognition using optical flow without complex feature extraction," *WSEAS Transactions on Computers*, vol. 6, pp. 763–770, 2007.

[9] S. Liao, W. Fan, A.C.S. Chung, and D.-Y. Yeung, "Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features," in *Proc. IEEE Conf. on Image Processing*, oct. 2006, pp. 665 –668.

[10] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.