Automatic Facial Expression Recognition on a Single 3D Face by Exploring Shape Deformation

Boqing Gong[1] Yueming Wang[1]* Jianzhuang Liu[1,2] Xiaoou Tang[1,2]
[1]Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2]Multimedia Lab, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
{gbq008, ymwang, jzliu, xtang}@ie.cuhk.edu.hk

ABSTRACT

Facial expression recognition has many applications in multimedia processing and the development of 3D data acquisition techniques makes it possible to identify expressions using 3D shape information. In this paper, we propose an automatic facial expression recognition approach based on a single 3D face. The shape of an expressional 3D face is approximated as the sum of two parts, a basic facial shape component (BFSC) and an expressional shape component (ESC). The BFSC represents the basic face structure and neutral-style shape and the ESC contains shape changes caused by facial expressions. To separate the BFSC and ESC, our method firstly builds a reference face for each input 3D non-neutral face by a learning method, which well represents the basic facial shape. Then, based on the BFSC and the original expressional face, a facial expression descriptor is designed. The surface depth changes are considered in the descriptor. Finally, the descriptor is input into an SVM to recognize the expression. Unlike previous methods which recognize a facial expression with the help of manually labeled key points and/or a neutral face, our method works on a single 3D face without any manual assistance. Extensive experiments are carried out on the BU-3DFE database and comparisons with existing methods are conducted. The experimental results show the effectiveness of our method.

Categories and Subject Descriptors

I.4.7 [**Image Processing and Computer Vision**]: Scene Analysis—*Shape*; I.4.8 [**Image Processing and Computer Vision**]: Feature Measurement—*Feature representation*

General Terms

Algorithms, Design, Experimentation

1. INTRODUCTION

Automatic facial expression recognition is an important research topic with many applications. In human-computer interface (HCI) community, affective computing employs human emotion to build more flexible and natural multimodal systems [5]. In face recognition, researchers have to pay great attention to handling the effect of expressions [10]. In 2D or 3D face retrieval, automatic facial expression recognition can serve as a particular kind of feature or a re-ranking algorithm to provide more accurate retrieval results. Moreover, 3D face models have been one of the five tracks of 3D Shape Retrieval Contest (SHREC) since 2007 [1].

In the past two decades, much effort has been paid on 2D expression recognition [3][7]. However, since 2D facial images are essentially projections of 3D human faces, facial expression recognition techniques based on them suffer from pose and illumination variations. With the rapid development and the dropping cost of 3D digital acquisition devices, 3D face data, which represent faces as 3D point sets or range data, can be captured more quickly and accurately. Several public 3D face databases have been available now [16][8]. 3D face data contain explicit 3D geometry, so more clues can be used to encode data changes caused by expressions and handle the variations of face poses. Thus, the use of 3D information in facial expression recognition has attracted attention and some techniques have been presented in recent years [14][11][13].

Based on the observation that 3D surface features represent intrinsic facial surface structures associated with specific facial expressions, Wang et al. [14] proposed a primitive surface feature stemming from two surface geometric features, curvature and gradient. They partitioned a 3D face into seven regions guided by the neuro-anatomy knowledge and obtained the statistical primitive feature distribution in each region. They showed that their algorithm is better than two 2D appearance based methods using Gabor wavelets and topographic context. Note that their partitioned regions do not contain the mouth and eyes, which are often used as expressional regions in 2D image based methods [6], and their algorithm involves manually labeled feature points in order to obtain more accurate region partitions.

Soyel and Demirel [11] represented different facial actions by six characteristic distances using eleven manually labeled feature points. They compared their 3D distance vector based facial expression recognition (3D-DVFER) algorithm to the 2D appearance based Gabor-wavelet algorithm. The experiments showed the superiority of their 3D-DVFER. Tang and Huang [13] further explored the effect of the distance vectors using more manually labeled feature points. They presented both manually designed and automatically selected distance vectors using a feature selection algorithm based on Kullback-Leibler divergence. To achieve the person-independent requirement, Soyel and Demirel [11] normalized the distance vector of an expressional face by the width of the face, while Tang and Huang [13] normalized the distance vector of an expressional face by facial animation parameter units (FA-

*Corresponding author.

MM'09, October 19–24, 2009, Beijing, China.

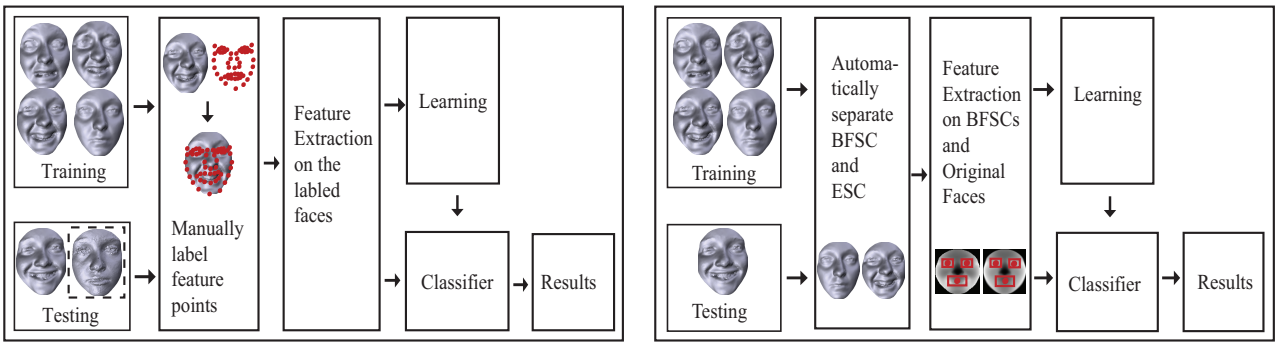

Figure 1: Comparison between the previous framework and our framework. (a) The framework of the previous methods. (b) The framework of our method.

PUs) in its corresponding neutral face as guided by the MPEG-4 standard.

The above methods have demonstrated the superiority of 3D face based methods over 2D image based ones. The main drawbacks of these methods are that they all need manually labeled facial key points for facial expression recognition in both training and testing processes. In addition, [13] also needs the help of neutral faces. These drawbacks make these methods can only work under some constrained conditions.

This paper proposes an automatic facial expression recognition approach based on a single 3D face. An expressional 3D facial surface is assumed as an approximate sum of two parts. One is a basic facial shape structure which contains little information of expressions and is commonly person-specific. The other is expressional shape component (ESC) which includes rich information about expressions. The ESC is expression-specific, and thus ESCs caused by similar expressions are also similar among a large range of different facial samples. The basic facial shape component (BFSC) is estimated from a group of aligned training data and the input expressional face. After that, based on expressional regions of the BFSC and the original expressional face, the shape depth information is encoded as expression descriptors which are used for a Support Vector Machine (SVM) for classification. The whole framework of our method is shown in Fig. 1(b), together with previous one for comparison in Fig. 1(a). It not only performs better than previous ones, but also is independent of the manual labeling of facial feature points. To the best of our knowledge, this is the first 3D facial expression recognition algorithm without the need of the manually labeling process.

2. SEPARATION OF BFSC AND ESC

The ESC of an expressional 3D face is the surface deformation of the basic face shape, e.g., the neutral face. Since neutral faces are not always available in expression recognition problems, we propose a learning-based method to estimate the basic face shape of an input expressional 3D face. The estimation uses the information of a group of neutral faces and the input expressional face. Then, the ESC is separated by subtracting the basic face shape from the original expressional face. In addition, before estimations of the basic faces, all 3D face samples are put in a standard coordinate system by an automatic alignment method.

2.1 3D Face Alignment

We use the same preprocessing as Wang et al.'s in [15] to align every face in a standard coordinate system. Let S denote the point set of a 3D face, and S^m be its mirror set with respect to some plane E_m. Then the ICP algorithm [2] is adopted to register S^m to S. A facial symmetry plane can be defined as the fitting plane of the midpoint set $B = \{p_i^b | p_i^b = (p_i + p_i^{m\prime})/2, p_i \in S, p_i^{m\prime} \in S^{m\prime}\}$, where $S^{m\prime}$ is the corresponding point set of S^m after the registration. Based on this symmetry plane, the central profile is found and two key points, the nose tip and the top of nose bridge, are robustly extracted on the profile. we can define a standard coordinate frame, in which the origin is the nose tip and the three axes are placed as shown in Fig. 2. The detail can be found in [15]. After the alignment, a 3D face can be represented by a depth image, which is obtained by sampling the projection of the 3D face on the x-y plane with a size of 100×100 in our experiments.

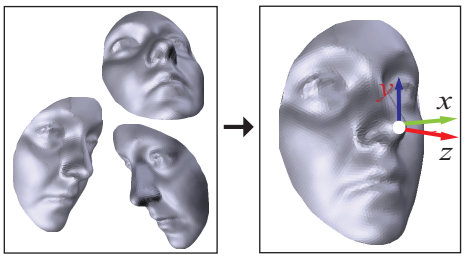

Figure 2: 3D face alignment in the standard coordinate frame.

This process allows us to estimate the basic face structure from its corresponding expressional face placing in a standard coordinate frame, which is described in the next subsection. Furthermore, it makes our method without the need of manually labeling the feature points on the faces.

2.2 Estimation of BFSC

Our BFSC estimation is based on the assumption that given sufficient training samples, a new face can be recovered approximately by the linear combination of the training faces. Suppose that each depth image is represented by a vector $\mathbf{x} \in \mathbb{R}^N$ and there are M training samples $\{\mathbf{x}_i\}_{i=1}^M$ which are all neutral faces. Then we approximate the basic face $\hat{\mathbf{x}}_e$ of an expressional face $\mathbf{x}_e$ by the linear combination of the training samples:

$$\hat{\mathbf{x}}_e \approx \sum_i c_i \mathbf{x}_i = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M][c_1, c_2, \cdots, c_M]^T = X\mathbf{c}, \tag{1}$$

where $\mathbf{c} = [c_1, c_2, \cdots, c_M]^T$ and $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M]$.

One possible solution to determining the weights $\{c_i\}_{i=1}^M$ is to use the discrete Karhunen-Loeve transform (KLT) [9]. Without loss

of generality, suppose that $\bar{\mathbf{x}} = \sum_i \mathbf{x}_i = 0$[1]. Let $\{(\mathbf{e}_i, \lambda_i), i = 1, 2, \cdots, N\}$ be the eigensystem of XX^T, where $\lambda_1 \geq \lambda_2 \geq \cdots, \geq \lambda_N$. Also let $P = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N]$. These eigenvectors span a linear face space. Many eigenvectors are devoted to individual differences in face structure, while noise, mainly facial expressions here, are represented orthogonal to these eigenvectors [6]. So we can reconstruct the neutral face structure if proper eigenvectors are selected, which are often the first K ($K \leq M$) eigenvectors. Let $\hat{P} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_K]$. Thus, the reconstructed basic face can be approximately represented as

$$\hat{\mathbf{x}}_e \approx \hat{P}\mathbf{w}, \tag{2}$$

where $\mathbf{w} = \hat{P}^T\mathbf{x}_e$. Note that $\mathbf{x}_e$ is the expressional face being recognized while $\hat{\mathbf{x}}_e$ is its corresponding basic face being reconstructed by the projections of $\mathbf{x}_e$ on the selected eigenvectors. The matrix P can be computed from $P = XVQ$, where V is the matrix formed by the eigenvectors of X^TX, $Q = [D|O]$, $D = diag(\sqrt{\lambda_1^{-1}}, \sqrt{\lambda_2^{-1}}, \cdots, \sqrt{\lambda_M^{-1}})$ is a diagonal matrix, and O is an $M \times (N - M)$ zero matrix [4]. So,

$$\hat{\mathbf{x}}_e \approx \hat{P}\mathbf{w} = X\hat{V}\hat{Q}\mathbf{w}, \tag{3}$$

where $\hat{V}$ is formed by the first K columns of V and

$$\hat{Q} = diag(\sqrt{\lambda_1^{-1}}, \sqrt{\lambda_2^{-1}}, \cdots, \sqrt{\lambda_K^{-1}}). \tag{4}$$

With (3) and (1) we have $\mathbf{c} = \hat{V}\hat{Q}\mathbf{w}$. Finally, the BFSC $\hat{\mathbf{x}}_e$ of an expressional face is computed by $X\mathbf{c}$.

3. EXPRESSIONAL REGIONS AND AN EXPRESSION DESCRIPTOR

In expressional face images, the eye regions and mouth regions are considered as expressional areas containing rich information of expressions [6]. Since the depth images are built based on the aligned 3D faces, these regions can be easily extracted by a mask. As shown in Fig. 3(a), we randomly select 300 depth images from the training samples, and calculate an average depth image (see Fig. 3(b)). The centers of the eyes and mouth together with three rectangle regions are used as a mask for extracting the expressional regions. We find that the recognition results are not sensitive to the sizes of the rectangle regions. Thus, the size of each eye region is set to 32×20 and the size of the mouth region is set to 40×25, empirically. The main advantage of the mask is that it helps to find key regions without the need of manual labeling in the testing.

With the BFSC and the original expressional face, the deformation of facial surface can be captured by encoding the depth differences between them. By the expressional region mask, the gray levels of the depth images within the three regions in the BFSC and the original face are arranged with the same order to form two vectors. Then an expression descriptor of a 3D face is defined as:

$$\mathbf{F}(f_e, f_b) = \mathbf{F}(f_e) - \mathbf{F}(f_b), \tag{5}$$

where $\mathbf{F}(f_e)$ and $\mathbf{F}(f_b)$ are the vectors extracted from the original face and its corresponding BFSC, respectively. The feature vectors are used for training and testing by an SVM algorithm.

4. EXPERIMENTS

In this section, we test our facial expression recognition method on a database named BU-3DFE [16]. The database contains 100 subjects with both males and females and a variety of ethnic ancestries and ages. 25 faces are captured for each subject, i.e., 6 prototypical expressions each with 4 different intensities and a neutral face (see the detailed description of the database in [16]).

[1]This can be obtained by subtracting the mean of all the training samples from every sample.

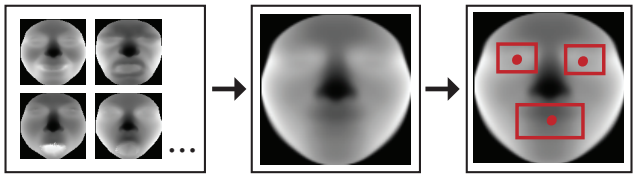

Figure 3: Expressional region mask. (a) A group of aligned depth images. (b) The average depth image. (c) The mask including the eye regions and the mouth region.

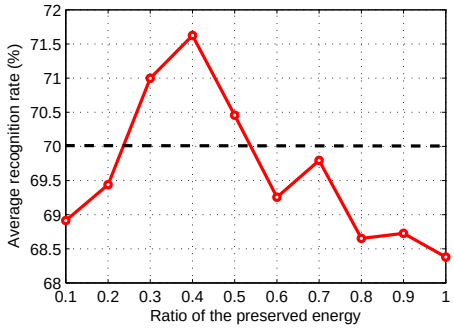

Figure 4: Average recognition rate vs. the ratio of preserved energy in the BFSC estimation.

The BU-3DFE database is used in existing facial expression recognition work [14][11][13]. The previous common setting is that 720 3D faces of 60 subjects are selected in their experiments, each subject with 12 expressional samples (2 higher intensities for every kind of expression). The 60 subjects are randomly partitioned into two subsets, a training set with 54 subjects (648 samples) and a test set with 6 subjects (72 samples). According to different partitions (54 vs. 6), 20 independent experiments are performed in [14], while in [11] and [13], 10 experiments are conducted. The reported results are the averages of the results of the independent experiments. In our experiments, we use the similar setting for comparison purpose.

4.1 Testing the Ratio of Preserved Energy in the BFSC Estimation

This experiment tests how different ratios of preserved energy in the BFSC estimation affect the recognition. Ten ratios from 0.1 to 1.0 with an interval of 0.1 are tested and Fig. 4 shows the average recognition rates. The highest recognition rate of 71.63% is achieved at the energy ratio of 0.4. This ratio is then used for the subsequent experiments.

4.2 Comparison with Related Work

In this section, we compare our method with the related work using the 54-versus-6-subject partitions. As pointed out in the beginning of this section, different partitions are independently trained and tested, and the average of all the results is shown as the final result. The partition process should guarantee that every subject is tested at least once. This subject-based partition aims to test how well the algorithm is with respect to the person-independent requirement in facial expression recognition. The selected 60 subjects are the same as those in [14]. However, we find that neither 20 [14] nor 10 [11][13] times of experiments are enough to have a sta-

Table 1: The comparison of different facial expression descriptors.

%	$\mathbf{F}(f_e)$	$\mathbf{F}(f_e, f_n)$	$\mathbf{F}(f_e, f_b)$
dist-soyel [11]	67.52	–	–
dist-tang [13]	–	74.51	–
prim-curv [14]	61.79	–	–
ours (depth)	**68.77**	**76.22**	**71.63**

Table 2: The average confusion matrix obtained by our method.

%	AN	DI	FE	HA	SA	SU
AN	**71.41**	12.28	2.92	0.00	15.30	2.48
DI	9.87	**76.60**	7.18	2.03	2.84	2.94
FE	3.91	4.92	**62.48**	15.33	2.87	4.76
HA	0.72	2.43	9.32	**81.21**	0.00	0.49
SA	14.06	1.11	4.56	0.00	**77.49**	1.19
SU	0.03	2.66	13.52	1.42	1.50	**88.13**

ble result. The average recognition accuracy obtained by 10 or 20 random experiments varies greatly, from about 50% to more than 90%. So we run the experiments 1000 times independently and obtain stable average recognition accuracies for all the methods. All the results below are obtained using this experimental setting. The RBF kernel of SVM is used for classification in the four methods.

We compare different facial expression descriptions in Table 1, including Soyel and Demirel's distance vector (dist-soyel) [11], Tang and Huang's manually designed distance vector (dist-tang) [13], primitive surface feature (prim-curv) similar to [14] but obtained based on [12], and our depth feature (depth).

In Table 1, the second column records the recognition results of the features obtained from the expressional faces only ($\mathbf{F}(f_e)$), the third column shows the results of the features based on both the expressional faces and the neutral faces ($\mathbf{F}(f_e, f_n) = \mathbf{F}(f_e) - \mathbf{F}(f_n)$, where f_n denotes a neutral face), and the fourth column shows the result from the expressional faces and the estimated BFSCs ($\mathbf{F}(f_e, f_b)$). Obviously, $\mathbf{F}(f_e, f_b)$ performs better than $\mathbf{F}(f_e)$, which indicates that the separation of BFSC and ESC is effective for expression recognition. When compared with [13] that requires neutral faces and the manual labeling (the third column), our method (also using the neutral faces) still obtains better result. It should be noted that the previous three methods all need manually labeled facial key points for the recognition, while ours is automatic.

One common characteristic of all the descriptors is that they all recognize "Happiness" and "Surprise" better than other types of facial expressions. The average confusion matrix obtained by our algorithm is shown in Table 2, where AN, DI, FE, HA, SA and SU are short for "Anger", "Disgust", "Fear", "Happiness", "Sadness" and "Surprise", respectively.

5. CONCLUSION

In this paper, we have developed an automatic 3D facial expression recognition algorithm requiring no manual facial keypoint labeling assistance. An expressional face is separated as a basic facial shape component (BFSC) and an expressional shape component (ESC). The description of facial expressions is designed based on both the original expressional face and its BFSC. Our algorithm obtains the highest average recognition rates in the comparison experiments. In addition, we find that the neutral face plays an important role in improving the facial expression recognition accuracy, which further shows that the separation of BFSC and ESC should be an important part of a facial expression recognition system.

6. ACKNOWLEDGMENTS

This work was supported by grants from Microsoft Research Asia (FY09-RES-OPP-102), the Research Grants Council of the Hong Kong SAR, China (Project No. CUHK 414306, 415408), and Shenzhen Bureau of Science Technology & Information, China.

7. REFERENCES

[1] *SHREC: Shape Retrieval Contest*. http://www.aim-at-shape.net.

[2] P. J. Besl and N. D. McKay. A Method for Registration of 3-D Shapes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[3] B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36(1):259–275, 2003.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[5] A. Jaimes and N. Sebe. Multimodal Human–Computer Interaction: A Survey. *Computer Vision and Image Understanding*, 108(1-2):116–134, 2007.

[6] C. Padgett and G. Cottrell. Representing Face Images for Emotion Classification. *Advances in Neural Information Processing Systems*, pages 894–900, 1997.

[7] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[8] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 947–954, 2005.

[9] T. Russ, C. Boehnen, and T. Peters. 3D Face Recognition Using 3D Alignment for PCA. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, 2006.

[10] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE*, 94(11):1948, 2006.

[11] H. Soyel and H. Demirel. Facial Expression Recognition Using 3D Facial Feature Distances. In *Int'l Conf. on Image Analysis and Recognition*, pages 17–22, 2006.

[12] H. Tanaka, M. Ikeda, and H. Chiaki. Curvature-Based Face Surface Recognition Using Spherical Correlation. Principal Directions for Curved Object Recognition. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 372–377, 1998.

[13] H. Tang and T. S. Huang. 3D Facial Expression Recognition Based on Automatically Selected Features. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[14] J. Wang, L. Yin, X. Wei, and Y. Sun. 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 17–22, 2006.

[15] Y. Wang, X. Tang, J. Liu, G. Pan, and R. Xiao. 3D Face Recognition by Local Shape Difference Boosting. In *European Conference on Computer Vision*, 2008.

[16] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D Facial Expression Database for Facial Behavior Research. In *Int'l Conf. on Automatic Face and Gesture Recognition*, 2006.