# FACIAL EXPRESSION SYNTHESIS BASED ON MOTION PATTERNS LEARNED FROM FACE DATABASE

*Jia Jia, Shen Zhang, and Lianhong Cai*

State Key Laboratory on Intelligent Technology and Systems, National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University

## ABSTRACT

Facial expression is the core function in face-to-face human-computer communication. In order to improve the accuracy and variety of the synthesized facial expressions, we propose a facial expression synthesis approach based on motion patterns learned from face database. We first define a set of partial facial regions: eyebrow, eye-lid, eyeball, upper lip, bottom lip and corner lip. For each region, we use a hierarchical clustering algorithm to learn the motion patterns from the face database. Then the patterns are used for parameterized synthesis of facial expression on the target face models. The experimental results show the effectiveness of the proposed approach.

***Index Terms***— facial expression synthesis, motion pattern.

## 1. INTRODUCTION

Facial expression synthesis plays an important role in implementing intelligent interface of human computer communications. The study on facial expression synthesis has attracted researchers' interests for a long time [1]. Recent works put forward the concept of "facial expression cloning" [2], the process of which transfers motion vectors from a source face model to a target model. Several example-based approaches are proposed to synthesize facial expressions by facial expression cloning [3, 4]. Further investigations are devoted to the feature-based facial expression synthesis methods [5]. Different with the example-based approaches, these works establish the face model using different facial features, and a scattered feature interpolation technique is adopted for synthesis [6-8]. However, most previous works need to pre-define expression patterns for different emotions, such as happy, angry, surprise, and so on. And the whole face area is usually modeled as one pattern. These problems lead to less expressivity in facial expression synthesis.

To address the problems, we propose a novel facial expression synthesis approach based on the statistical analysis of facial motion patterns. We first define a set of partial facial regions, such as eyebrow, mouth and eye. Each partial facial region is represented by several MPEG-4 facial animation parameters (FAPs). For the FAPs of each partial facial region, we use a hierarchical clustering algorithm to learn the motion patterns from a public face database. And finally, the motion patterns are used for facial expression synthesis on different target faces. Compared to the traditional facial expression synthesis methods, our proposed approach has the following advantages:

1) previous works synthesize facial expressions based on pre-defined expression patterns. Instead, we use a clustering algorithm to learn the motion patterns from face database. With the learned various patterns, the synthesized expressions could be more rich and natural.
2) most previous studies treat the whole face area as one pattern. Facial expressions are usually composed of partial patterns, such as eyebrow-raise, mouth-bent, and eye-open. Therefore, we believe that analysis and modeling for different face organs can help improve the accuracy of facial expression synthesis.

We conduct an experiment, in which statistical clusters of motion patterns are obtained from a public face database, and then used for facial expression synthesis on different photographs and cartoon images. The synthesized facial expressions show the effectiveness of our proposed approach.

## 2. LEARNING FACIAL MOTION PATTERNS FROM FACE DATABASE

### 2.1. Definition of Partial Facial Regions

In this sub-section, we define a set of partial facial regions to describe the facial expressions. Different with previous studies which capture the whole face motion as a single pattern, we focus on three main facial organs: eyebrow, eye, and mouth. From the definition of MPEG-4 facial definition points (FDP), we selected the facial points for each organ

This work is supported by National Natural Science Foundation of China (60805008, 90820304), the National Basic Research Program of China (2006CB303101), and the National High Technology Research and Development Program of China (2009AA011905).

using the facial animation parameters (FAP). Table 1 illustrates the partial facial regions and their corresponding FAPs.

Table 1. Definition of partial facial expressions

| | Partial Facial Region | FAP Index Number |
|---|---|---|
| 1 | Eyebrow | 31, 32, (33), 34, 35, 36, 37, 38 |
| 2 | Eye-lid | 19, (20), 21, 22 |
| 3 | Eyeball | (23), 24, 25, 26 |
| 4 | Upper Lip | (4), 8, 9, 51, 55, 56 |
| 5 | Bottom Lip | 5, 10, 11, (52), 57, 58 |
| 6 | Corner Lip | (12), 13, 59, 60, 6, 7, 53, 54 |

### 2.2. Hierarchical Clustering of FAPs in Partial Facial Regions

The Cohn-Kanade facial expression database [9, 10] is used in our study to learn the motion patterns for each partial facial region. The database contains the image sequences of 97 university students, 65% of whom are females, performing 23 kinds of facial expressions, shown as Fig.1 (a). To parameterize the facial motions, we use the FAP annotation provided by LAIV lab [11] as shown in Fig.1 (b), covering all the facial points defined in Table 1. The FAPs are normalized first by the facial animation parameter unit (FAPU) to eliminate the personal difference caused by the size of face and facial organs.

(a) (b)

Fig.1. (a) Facial expression images (b) Feature points annotation

We use the hierarchical clustering algorithm to learn the different motion patterns for each partial facial region. The hierarchical clustering algorithm constructs a confusion tree based on comparing the same partial facial regions of facial expression samples. Different with other clustering methods, it is very flexible to determine the number of clusters by pruning the confusion trees. The motion patterns are learned by the similarity analysis between facial expression samples. Before clustering, we firstly define the similarity between each pair of FAPs as their Euclidean distance $d_{rs}$:

$$d_{rs} = \left( \sum_{n} \left[ FAP(r,n) - FAP(s,n) \right]^2 \right)^{\frac{1}{2}} \tag{1}$$

where *FAP(r, n)* and *FAP(s, n)* represent the *n*th FAP values for facial expression samples *r* and *s*. During clustering, the algorithm groups all the samples into a binary, hierarchical clustering tree. At initial step, all the samples are taken as leaf nodes. Pairs of leaf nodes that are close will be grouped into clusters based on their distance to each other. The newly formed clusters will further linked to others if they have small distances. The distance between two clusters *A* and *B* is defined as follows:

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij} \tag{2}$$

where $n_A$ (or $n_B$) represents the number of samples in *A*(or *B*).

The algorithm continues to create bigger clusters until all the facial expression samples in database are linked together in a hierarchical tree. For each link node in the hierarchical tree, its height is defined as the distance between its two sub-nodes. Then the pruning process is executed according to the height of each node. The pruning process keeps finding the node having the smallest height. And a horizontal cut will be executed until the tree only leaves $C_n$ clusters ($C_n$ is a threshold). To obtain an appropriate number of motion pattern clusters for each partial facial region, we use the *Silhouette Coefficient* to measure how appropriate each data sample is clustered [12]. The silhouette coefficient is calculated as follows:

$$\begin{aligned} & S_A(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}, \quad i \in A \\ & a(i) = \frac{1}{N_A - 1} \sum_{j \neq i, j \in A} d_{ij}, \quad b(i) = \min\left( \frac{1}{N_B} \sum_{j \in B} d_{ij} \right) \\ & \overline{S}_A = \frac{1}{N_A} \sum_{i=1}^{N_A} S_A(i) \end{aligned} \tag{3}$$

where the data sample *i* belongs to cluster *A*, $N_A$ is the number of samples in *A*, *a(i)* is the average similarity of *i* with all other samples within cluster *A*, *b(i)* is the lowest average similarity of *i* with the samples in another cluster ( for example, cluster *B*) which *i* is not a member of. $S_A(i)$ is the value of silhouette coefficient for *i*, which ranges from -1 (where *i* is appropriately clustered (in cluster *A*)) to +1 (where the *i* would be more appropriate if it was clustered in its neighboring cluster (in cluster *B*)). The average silhouette coefficient of all the data samples in cluster *A* ($\overline{S}_A$) is taken as a measure of how appropriate the data has been clustered in *A*. In this study, the final number of clusters is experimentally determined for each partial facial region. That means we change the $C_n$ from 2 to *n* (*n*=16 in our study), and choose the one which has the minimum average $\overline{S}$ as the final number of clusters.

The advantage of hierarchical clustering is that it keeps the characteristic of every data sample, which means that even the particular motion patterns with few data samples could be maintained in the clustering trees. By such hierarchical clustering scheme, we can learn all the possible facial motion patterns.

### 2.3. Learning Facial Motion Patterns based on FAPs Correlations

In the hierarchical clustering, the facial motions with the similar FAP values are grouped into the same cluster. In this sub-section, we proceed to learn the motion patterns based on the correlation analysis among the FAPs within a partial facial region.

The previous study by Lavagetto *et al*. [13] has proved that there exists high correlation among FAPs. For a partial facial region, we calculate the correlation matrix **R** of its FAPs, shown as Eq.4. $M$ represents the number of FAPs in this region. $\mathbf{FAP}_i$ and $\mathbf{FAP}_j$ represent the vectors composed of the $i$th and $j$th FAPs of all the data samples respectively ($i, j \in [1, M]$). $r_{ij}$ represents the *Pearson Correlation Coefficient* between $\mathbf{FAP}_i$ and $\mathbf{FAP}_j$. $\mu_i$ and $\mu_j$ represent the mean values of $\mathbf{FAP}_i$ and $\mathbf{FAP}_j$ respectively. $\mathrm{E}[\bullet]$ is the *Mathematical Expectation*. And **I** is the unit vector.

$$\mathbf{R} = \begin{bmatrix} r_{11} & \cdots & r_{1M} \\ \cdots & r_{ij} & \cdots \\ r_{M1} & \cdots & r_{MM} \end{bmatrix}, \quad r_{ij} = \left| \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \right| \tag{4}$$

$$C(i,j) = \mathrm{E}\left[ (\mathbf{FAP}_i - \mathbf{I} \cdot \mu_i)(\mathbf{FAP}_j - \mathbf{I} \cdot \mu_j) \right] \quad (i, j \in [1, M])$$

The sum of each row in **R** is calculated to select the $\mathbf{FAP}_{rep}$ which has the largest correlation with other FAPs, shown as Eq.5. For each partial facial region, the index number of $\mathbf{FAP}_{rep}$ is illustrated with parenthesis in Table 1. For a cluster of a partial facial region, the relation between each FAP in the $n$th sample ($n \in [1, N]$) and $FAP^n_{rep}$ ($FAP^n_{rep} \in \mathbf{FAP}_{rep}$) is described as Eq.6, where $N$ is the number of samples in this cluster. The other non-representative FAPs are interpolated by the $FAP^n_{rep}$ with coefficient $\alpha_i$, which is determined by the least square estimation, shown as Eq.7.

$$\mathbf{FAP}_{rep} = \mathbf{FAP}_k, \quad k = \arg\max_{i=1,\ldots M} \left( \sum_{j=1}^{M} r_{ij} \right) \tag{5}$$

$$FAP^n_i = \alpha_i \bullet FAP^n_{rep} \quad (i \neq k, n \in [1, N]) \tag{6}$$

$$\alpha_i = \frac{\sum_{n=1}^{N} FAP^n_i \cdot FAP^n_{rep}}{\sum_{n=1}^{N} (FAP^n_{rep})^2} \tag{7}$$

For each cluster, we calculate the $\mathbf{FAP}_{rep}$ and obtain the interpolation coefficient $\alpha_i$ for other FAPs. Suppose that there are $L$ clusters for the $t$th partial facial region which has $M_t$ FAPs. Each cluster represents a motion pattern. For each cluster, we obtain a vector of interpolation coefficient $\mathbf{V}^l_t$ as shown in Eq.8. We use the $\mathbf{V}^l_t$ to represent the $l$th motion pattern in the $t$th partial facial region. By such representation, we can parameterize the motion patterns, and it is convenient for facial expression synthesis with FAPs.

$$\mathbf{V}^l_t = [1, \alpha^l_2, \cdots, \alpha^l_{M_t}], l \in [1, L] \tag{8}$$

## 3. FACIAL EXPRESSION SYNTHESIS BASED ON MOTION PATTERNS

In this section, we propose a method for parameterized facial expression synthesis based on motion patterns.

For an input face image, a face alignment toolkit [14] is first used to locate 88 facial feature points, shown as Fig.2 (a and b). Based on the alignment points, a face mesh model [15] is fitted to the 88 facial points, as shown in Fig.2 (c). This personalized face model consists of 361 mesh grids in triangle. The design of the mesh is based on the facial organs location and the movement direction of facial muscles, which is convenient for parameterized facial expression animation.

(a) (b) (c)

Fig.2. Face model fitting on 2-D front face image

To synthesize facial expression, the value of $FAP_{rep}$ for each partial facial region can be manually set from one extreme state (such as mouth-open) to the opposite extreme state (such as mouth-close). And the values of other FAPs in the same partial facial region are interpolated according to the linear function that represents a specific motion pattern learned from the data clustering (Eq.6, Eq.8). The generated FAPs are then used to animate the face wireframe models.

## 4. EXPERIMENT

In this section, we conduct an experiment to obtain the statistical clusters of facial motion patterns. In order to demonstrate the universality of our proposed approach on different face models, we would like to show more synthetic facial expressions on photographs and cartoon images in Fig.4, based on the learned facial motion patterns.

Totally 486 facial expression images from the Cohn-Kanade facial expression database are selected as the training set. For each partial facial region, we build a hierarchical clustering tree based on the annotated FAPs. To obtain the proper clusters for each facial region, we take the number of clusters ($C_n$) as threshold to prune the hierarchical tree. To further examine the pruning results, we use the average silhouette coefficient to measure how appropriate the samples are clustered. Fig.3 illustrates the average silhouette coefficients with different threshold $C_n$. To obtain the final division of facial motion patterns, we take the cluster number where the average silhouette coefficient gets local minimum and starts to increase

continually. Table 2 shows the final clusters for each partial facial region.

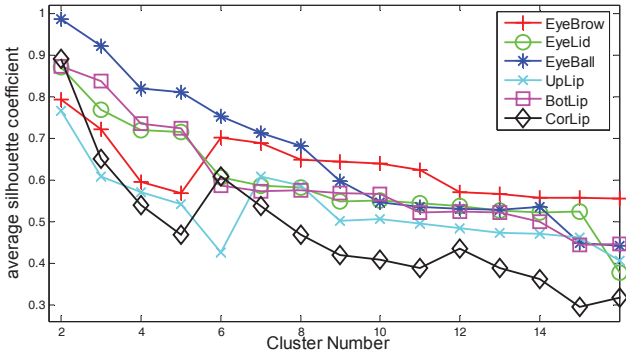


Fig.3. Cluster numbers and the average silhouette coefficients

Table 2. Statistics of facial motion patterns

| Partial Facial Region | Eye brow | Eye lid | Eye ball | Upper Lip | Bottom Lip | Corner Lip |
|---|---|---|---|---|---|---|
| Cluster num | 5 | 6 | 10 | 6 | 6 | 5 |
| Average Silhouette | 0.57 | 0.61 | 0.55 | 0.42 | 0.58 | 0.47 |

We apply the learned motion patterns for facial expression synthesis on different facial images, such as a real human photo [Fig.4 (a)], a cartoon character [Fig.4 (b)], and also a stick figure generated from real photo [Fig.4 (c)]. The synthesized results are shown as Fig.4 (1)-(9). The images in the same columns (such as (1), (4), (7)) use the same motion patterns to generate the corresponding facial expressions. The synthetic facial expression image has shown the effectiveness and expressiveness of our approach.

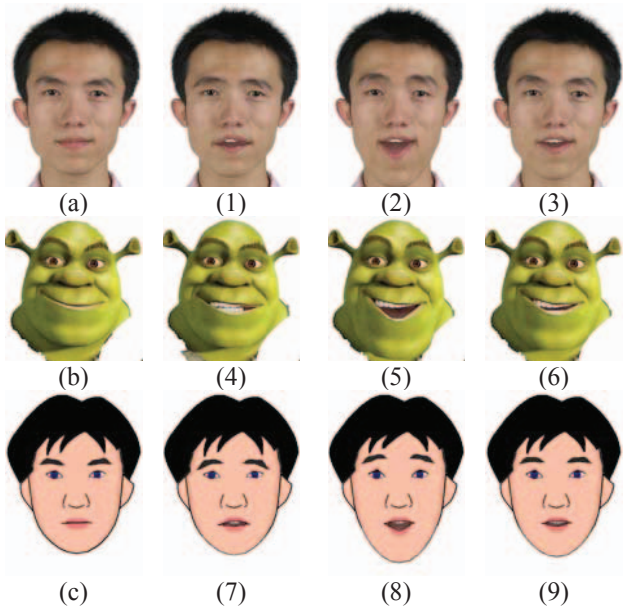


Fig.4 Synthesized facial expressions based on different motion patterns

## 5. CONCLUSIONS

In this paper, we present a facial expression synthesis approach based on motion patterns which are learned from face database. The synthesized facial expressions on different face models demonstrate the effectiveness and expressiveness of our approach.

## 6. REFERENCES


[1] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Tran.PAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.

[2] J.Y.Noh and U.Neumann, "Expression cloning" in *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, Los Angeles, 2001, pp. 277 – 288.

[3] H.Pyun, Y.Kim, W.Chae, H.W.Kang, *et al.*, "An example-based approach for facial expression cloning," in *Proceedings of Eurographics/SIGGRAPH Symposium on Computer Animation*. New York, NY, USA: ACM, 2003, p. 23.

[4] P.W.Hsu, Y.Chang, C.K.Hsieh, *et al.*, "Facial expression cloning: using expressive ratio image and FAP to texture conversion," in *Proceedings of the 4th International Conference on Information, Communications and Signal Processing*, 2003.

[5] B. Park, H. Chung, T. Nishita, and S. Y. Shin, "A feature-based approach to facial expression cloning," *Computer Animation and Virtual Worlds*, vol. 16, pp. 291–303, 2005.

[6] S. Krinidis and I. Pitas, "Facial expression synthesis through facial expressions statistical analysis," in *Proceedings of 14th European Signal Processing Conference*, 2006.

[7] S.Kshirsagar, C.Joslin, W.S.Lee, *et al.* "Personalized face and speech communication over the internet," in *Proceedings of IEEE,Virtual Reality*, 2001.

[8] N.P. Chandrasiri, T. Naemura, and H. Harashima, "Interactive analysis and synthesis of facial expressions based on personal facial expression space," in *Proceedings of the Sixth IEEE International Conference on AFGR*, 2004.

[9] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of Fourth IEEE International Conference on AFGR*, Mar. 28–30, 2000, pp. 46–53.

[10] B.Theobald, I.Matthews, J.Cohn, and S. Boker, "Real-time expression cloning using active appearance models", in *Proceedings of* in *Proceedings of the ACM International Conference on Multimodal Interfaces*, 2007.

[11] G. Lipori. Manual annotations of facial fiducial points on the cohn kanade database. [Online]. Available: http://lipori.dsi.unimi.it/download/gt2.html

[12] Kaufman L., and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, 1990.

[13] F.Lavagetto and R. Pockaj, "An efficient use of MPEG-4 FAP interpolation for facial animation at 70 bits/frame," vol. 11, no. 10, pp. 1085–1097, 2001.

[14] L.Zhang, H.Ai, S.Xin, *et al.* "Robust face alignment based on local texture classifiers," in *The IEEE ICIP* 2005.

[15] Z.M.Wang, "Research on chinese viseme modeling and visual speech," Ph.D. dissertation, Tsinghua University, 2002.