Modeling the Correlation between Modality Semantics and Facial Expressions

Jia Jia[*], Xiaohui Wang[*], Zhiyong Wu[#,†], Lianhong Cai[*,#] and Helen Meng[†,#]
* Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
† Human-Computer Communications Laboratory, Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
Contact E-mail: [#] zywu@sz.tsinghua.edu.cn, [*] jjia@tsinghua.edu.cn

***Abstract*—Facial expression plays an important role in face-to-face human-computer communication. Although considerable efforts have been made to enable computers to speak like human beings, how to express the rich semantic information through facial expression still remains a challenging problem. In this paper, we use the concept of "modality" to describe the semantic information which is related to the mood, attitude and intention. We propose a novel parametric mapping model to quantitatively characterize the non-verbal modality semantics for facial expression animation. In particular, seven-dimensional semantic parameters (SP) are first defined to describe the modality information. Then, a set of motion patterns represented with Key FAP (KFAP) is used to explore the correlations of MPEG-4 facial animation parameters (FAP). The SP-KFAP mapping model is trained with the linear regression algorithm (AMMSE) and an artificial neural network (ANN) respectively. Empirical analysis on a public facial image dataset verifies the strong correlation between the SP and KFAP. We further apply the mapping model to two different applications: facial expression synthesis and modality semantics detection from facial images. Both objective and subjective experimental results on the public datasets show the effectiveness of the proposed model. The results also indicate that the ANN method can significantly improve the prediction accuracies in both applications.**

I. INTRODUCTION

Facial expression is an important channel of human-human interactions. Statistics show that non-verbal gestures account for 55% of the meaning about feelings and attitudes that a speaker delivers, while word and voice only account for 7% and 38% respectively [1]. To enable computers to communicate like human beings, synthesizing and recognizing facial expressions are becoming popular research topics and have attracted interests from both academic and industrial communities.

Most state-of-the-art studies have put focus on synthesizing or recognizing facial expressions by considering several emotion categories [4], such as happy, angry, surprise, etc. However, in addition to emotions, people also express their moods, attitudes, intentions and even beliefs with their facial expressions. The subtle and complex changes of facial expressions usually convey abundant semantic information, including not only emotions but also communicative purposes and action readiness [2][3]. Hence, there is a clear need for methods and techniques to analyze and model the correlation between semantic information and facial expressions. Unfortunately, the issue of how to convey the rich semantic information with the facial expressions is often ignored. Linguists and psychologists use the concept of "modality" to refer to "the manner of speaking by which the speaker shows his/her attitude and position in the current conversation" [5]. Inspired by linguistics and psychology, the *modality semantic* in this paper is defined as the semantic information which is related to the subjective mood, attitude and intention. Several studies in psychology have given formal description of the semantic meaning of facial expression. For example, Russell proposed a dimensional cognitive model emphasizing the important role of contextual semantics in facial expressions [6]. However, without a quantitative measurement, the results of these researches cannot be directly applied to build facial models for computers.

In this paper, we focus on modeling the correlation between modality semantics and facial expressions. The problem is non-trivial and poses a set of unique challenges: 1) we need a measure to quantify the modality semantics conveyed by the facial expressions; 2) it is unclear how one should model the correlation between modality semantics and facial expressions; 3) it is important to apply the model to real applications to verify its effectiveness. Figure 1 illustrates the overview of the correlations between modality semantics and facial expressions. As shown by the top-left part of the figure, the modality semantics conveyed by a facial expression are parameterized by a set of semantic dimensions, including pleasure, strength, confidence, attention, nervousness, activation, and dominance. These dimensions can describe the modality semantics that are related to not only subjective mood (with pleasure and nervousness) but also attitude (with confidence, attention and activation) and intention (with strength and dominance). Each dimension of modality semantics can take different levels of quantitative values. For example, as shown in the figure, *pleasure* and *confidence* take the high level positive values; *strength*, *attention*, *activation* and *dominance* take the middle level positive values; while *nervousness* takes the high level negative value. Details on the parameterization of modality semantics will be elaborated in Section III.A. The top-right part of the figure shows several different facial expressions that are related to different modality semantics. We can easily see that the smile expression best matches the modality semantic as parameterized by the above semantic dimensions, while the other facial expressions have quite different (even opposite) meanings. Particularly, the shape of mouth, eyes and eyebrows are the key factors to convey different modality semantics with facial expressions. Different shapes and motions of facial organs, such

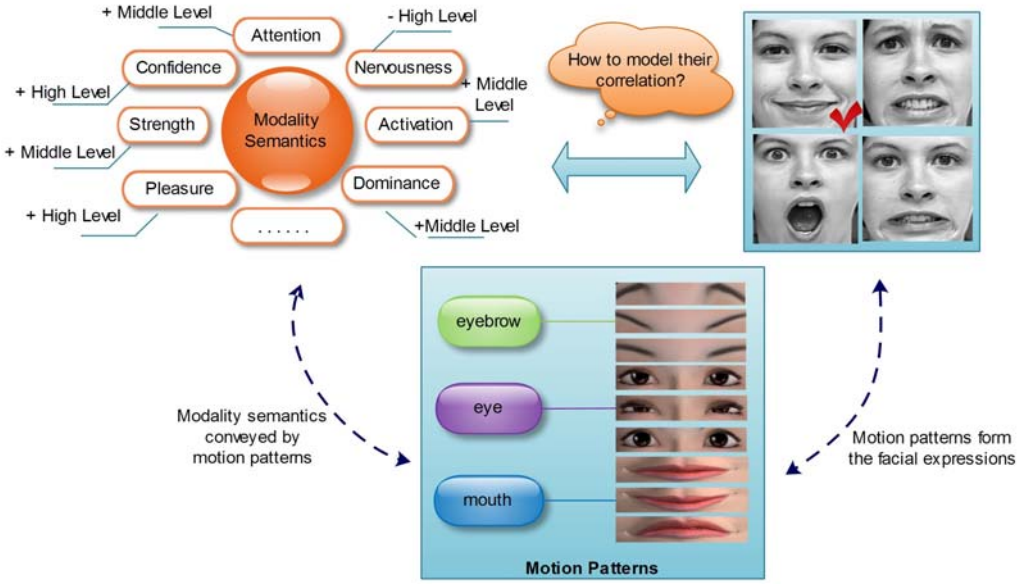


Fig.1. Illustration on modeling the correlation between modality semantics and facial expressions.

as raise eyebrows, squeeze eyebrows, open eyes, bend mouth, stretch mouth and so on are called motion patterns, as shown in the bottom part of the figure. Based on these observations, we propose our solutions for the above challenges in modeling the correlations between modality semantics and facial expressions: we first define a seven-dimensional semantic parameter (SP) to depict and quantify the modality semantics. Then based on MPEG-4 facial animation parameters (FAP), a set of facial expression parameters are selected to describe motion patterns. The selected FAPs are called Key FAPs (KFAP). We further use both linear and non-linear methods to model the correlation between SP and KFAP. Finally, the proposed models are applied in two applications to verify the effectiveness: facial expression synthesis driven by SP, and modality semantics detection from facial images.

The rest of this paper is organized as follows: Section II introduces related works. Section III formally defines SP and KFAP for parameterizing the modality semantics and facial expressions respectively. Section IV describes the proposed mapping model between SP and KFAP, using both linear and non-linear methods. Section V presents two applications using the proposed mapping model. Experimental results in section VI validate the rationality and effectiveness of our method. Finally, Section VII concludes.

II. Related Works

A. Facial Expression Analysis

Facial expressions and their relationships with emotions have been extensively studied [2]. Existing models for emotion description can be summarized into two categories: the categorical model and the dimensional model. For the categorical model, [7]-[10] were devoted to synthesizing or recognizing facial expressions for basic emotion categories (e.g. happy, sad, angry, etc.). However, the vagueness of natural language makes it hard to clearly describe the subtle variation of spontaneous emotion. It is also difficult to describe the continuum or non-extreme emotional states by discrete list of categories.

For the dimensional model, emotional states are quantitatively measured in terms of a small number of basic dimensions. Quantitative emotional analysis based on dimensional model is suitable for computer modeling [11]. Jia *et al.* [12] proposed an audio-visual speech synthesis approach for a Chinese avatar based on the pleasure-arousal-dominance (PAD) model.

However, all these methods can only deliver emotion information. In addition to emotions, people also express their moods, attitudes and intentions with their facial expressions, which are called "modality" by linguists and psychologists [5]. The analysis of how to convey the rich modality semantics besides emotion by facial expressions is often ignored by existing methods.

In this paper, we focus on the analysis of the rich modality semantics related to subjective mood, attitude and intention, and also the analysis of the correlation between modality semantics and facial expressions. A seven-dimensional semantic parameter (SP) is proposed to quantitatively describe modality semantics. This work extends existing work in two ways: 1) we enrich the meaning of facial expressions, from emotions to modality semantics; 2) we adopt and extend the dimensional approach to describe and quantify the modality semantics.

B. Facial Expression Representation

Most research on parametric facial expression representation [13] are based on the facial action coding system (FACS) proposed by Ekman, and the facial animation framework in MPEG-4 [14].

FACS is a human-observer-based system, which is designed to capture the subtle movements of isolated facial features. The basic facial motion patterns in FACS are called "action units" (AU). From the perspective of computer animation, the action units define the motion patterns of facial organs, but do not

provide the quantitative information required by facial animation. Therefore, the FACS system is widely adopted in the field of facial expression recognition, but it is not suitable for synthesis and animation purposes.

Contrary to FACS, the facial animation framework developed under MPEG-4 standards is designed completely for computer animation. In accordance to action units, the facial animation parameters (FAPs) are designed by the study of minimal perceptible muscle actions. There are in total 68 FAPs defined for quantifying movements of specific facial points. However, it is redundant to generate facial expressions by manipulating all the FAPs directly, since there is strong correlation between FAPs within the same facial organs. Such correlation in FAPs has been utilized for the coding of real time facial animation [15]. Controlling all the FAPs to deform the mesh is much more complex than using meaningful facial motion patterns.

In this paper, we use the motion patterns represented by the selected key FAPs (KFAP) to depict the movements of facial organs (eyebrow, eye, mouth), which combines the advantages of both human perceptible description method (FACS) and low-level deformation of face model (FAP).

III. Parameterization of Modality Semantics and Facial Expressions

In this section, we first present a formal definition of the modality semantic parameters (SP), and then introduce facial expression parameters represented with KFAP which is used to describe motion patterns of facial organs.

A. Parameterization of Modality Semantics

In psychological research, emotions have been studied for several decades, and recently the dimensional approach has gained popularity which can describe various human emotional states. Empirical study has been conducted to determine three essential dimensions to measure human emotions [22], namely *pleasure*, *arousal* and *dominance*. The dimensional description, suitable for computer modeling, captures the essential properties of emotion.

As has been described, in addition to emotions, people also express moods, attitudes and intentions with facial expressions. We adopt and extend the dimensional approach to describe such rich modality semantics conveyed by facial expressions [16]. Based on the semantic features proposed in componential analysis of semantics, seven dimensions are defined as shown in Table 1. Each dimension describes a different aspect of modality semantics. For example, *nervousness* and *pleasure* describe the subjective mood; *confidence, attention* and *activation* describe the subjective attitude; while *strength* and *dominance* describe the subjective intention. Table 1 presents these dimensions and their corresponding definitions.

The value of each dimension ranges from -1 to +1 (shown in Table 2), corresponding to the continuous changes from negative to positive, e.g. from "suspicion" to "confidence", or from "sadness" to "happiness". Zero values indicate that the corresponding dimension does not apply. According to the above definitions, we can describe a certain kind of modality semantics conveyed by a facial expression using a seven-dimensional semantic parameter (SP), which can be represented by a vector **SP** consisting of the values of each of the seven semantic dimensions. When we need to quantify the seven dimensions in our applications, the perceptual experiment will be conducted to annotate values for facial expressions in terms of the semantic dimensions, using a 5-point Likert scale, as shown in Table 2.

Table 1. Definition of dimensions of modality semantics

#	Dimension	Definition
1	Confidence	Belief towards events/situations
2	Strength	Influence on environment and people
3	Activation	Physical activity and mental alertness
4	Dominance	Status in inter-personal communication
5	Nervousness	Control of one's own emotional state/behavior
6	Pleasure	Positive/negative quality of emotional state
7	Attention	Interests in environment/events/people

Table 2. Annotation scales of modality semantic parameters (SP)

Score	Description
1	Obviously showing the positive state described by the semantic dimension
0.5	Generally showing the positive state described by the semantic dimension
0	Showing nothing that is related to the semantic dimension
-0.5	Generally showing the negative state described by the semantic dimension
-1	Obviously showing the negative state described by the semantic dimension

The dimensional approach is suitable for the description of modality, because of 1) the multi-dimensional property of modality; 2) the vagueness of modality; 3) the diversified expression of modality. Firstly, modality is a complex semantic category, which contains multiple dimensions. Secondly, the definitions of modality concepts are not absolutely determined. For example, unlike the concept of "gender" ("male" and "female") which has clear distinctions, there is no clear line between the modality concept of "good" and "bad". The modality concepts may be more suitably described by continuous dimensions. Thirdly, the modality can be expressed through various channels (e.g. text, images, speech, facial expressions, gestures, etc.). Different channels have their own characteristics, which lead to the diversified properties in cognition and expression. Different set of modality dimensions can meet the application needs in different scenarios. By dimensional description of modality, we aim to provide a quantitative model to achieve semantic computation.

It should be noted that, the above modality dimensions are designed for the facial expressions, which provides an instrument for semantic analysis and computation. However, these dimensions do not cover all the possible dimensions for modality. Based on the dimensional approach, different studies can define their own dimensions according to the specific research needs.

B. Parameterization of Facial Expressions

For parameterized analysis of facial expressions, the MPEG-4 facial animation parameters (FAPs) are adopted. However, since the FAPs focus on the control of a single facial point for face

model animation (e.g. raise the midpoint of the bottom outer lip), rather than the description of motion patterns of facial organs (e.g. a smiling mouth), it is complicated and not intuitive to define every FAP for a facial expression. Besides, previous studies have proved that strong correlation exists between FAPs within the same facial organs. For example, such correlation in FAPs was utilized for the coding of real time facial animation [15]. Inspired by [15], we define KFAP to describe the common motion patterns of eyebrow, eye and mouth regions.

We first define Motion Pattern (MP) as a subset of FAPs in the MPEG-4 standard. FAPs in one MP would be related to a certain motion of a facial organ, such as raise eyebrows, open eyes, etc. There are 9 MPs defined to represent the facial motions which contribute greatly in facial expressions. The definition of each MP is shown in Table 3. For example, $MP^{(2)}=\{FAP_{37}, FAP_{38}\}$ represent the motion pattern of squeeze eyebrows, and $MP^{(2)}_1$ is FAP_{37}. The MPs defined here are all mutually disjoint.

Table 3. Definition of motion patterns

#	Motion Pattern	FAPs in Each Motion Pattern (KFAP in the parenthesis)
1	Raise Eyebrows	FAP_{31}, FAP_{32}, (FAP_{33}), FAP_{34}, FAP_{35}, FAP_{36}
2	Squeeze Eyebrows	(FAP_{37}), FAP_{38}
3	Open Eyes	FAP_{19}, (FAP_{20}), FAP_{21}, FAP_{22}
4	Look Left/Right	(FAP_{23}), FAP_{24}
5	Look Up/Down	(FAP_{25}), FAP_{26}
6	Open Mouth (upper lip)	(FAP_4), FAP_8, FAP_9, FAP_{51}, FAP_{55}, FAP_{56}
7	Open Mouth (bottom lip)	FAP_3, FAP_5, FAP_{10}, FAP_{11}, (FAP_{52}), FAP_{57}, FAP_{58}
8	Bend Mouth	(FAP_{12}), FAP_{13}, FAP_{59}, FAP_{60}
9	Stretch Mouth	(FAP_6), FAP_7, FAP_{53}, FAP_{54}

A KFAP is the FAP selected from a MP as the one having the strongest correlation with the other FAPs in this MP. The KFAP is a representative FAP in a MP, and the values of the other FAPs in the MP can be derived from the value of the corresponding KFAP (Eq.5-Eq.6). In this way, a certain facial expression can be represented by a 9-dimension vector **KFAP** which consists of the value of these 9 KFAPs.

The KFAPs are selected with the following process:

For MP_i: firstly, we calculate the correlation matrix $\mathbf{R}_i$ for the FAPs in MP_i.

$$\mathbf{R}_i = \begin{bmatrix} r_{11}^{(i)} & \cdots & r_{1M_i}^{(i)} \\ \vdots & r_{uv}^{(i)} & \vdots \\ r_{M_i1}^{(i)} & \cdots & r_{M_iM_i}^{(i)} \end{bmatrix} \quad \text{where} \quad r_{uv}^{(i)} = \left| \frac{C_{uv}^{(i)}}{\sigma_u^{(i)} \sigma_v^{(i)}} \right| \quad 1 \le u, v \le M_i \tag{1}$$

$$C_{uv}^{(i)} = E(MP_u^{(i)} \cdot MP_v^{(i)}) - E(MP_u^{(i)}) \cdot E(MP_v^{(i)}) \tag{2}$$

$$KFAP_i \equiv MP_{key_i}^{(i)} \quad \text{where} \quad key_i = \arg\max_{u=1,2,\ldots M_i} (\sum_{v=1}^{M_i} r_{uv}^{(i)}) \tag{3}$$

where $r_{uv}^{(i)}$ is the absolute value of *Pearson's correlation coefficient* between $MP_u^{(i)}$ and $MP_v^{(i)}$. $C_{uv}^{(i)}$ is the covariance of $MP_u^{(i)}$ and $MP_v^{(i)}$, $\sigma_u^{(i)}$ and $\sigma_v^{(i)}$ are the standard deviations of $MP_u^{(i)}$ and $MP_v^{(i)}$ respectively. M_i is the number of FAPs in $MP^{(i)}$.

Then, the sum of $r_{uv}^{(i)}$ is calculated for each row of $\mathbf{R}_i$. The row with the largest sum is selected, and the *FAP* corresponding to this row ($MP_{key_i}^{(i)}$) is selected as KFAP for MP_i. The FAPs with parenthesis in Table 3 are the selected KFAPs for each MP. Thus, the **KFAP** is:

$$\begin{aligned} \mathbf{KFAP} &= \left(MP_3^{(1)}, MP_1^{(2)}, MP_2^{(3)}, MP_1^{(4)}, MP_1^{(5)}, MP_1^{(6)}, MP_5^{(7)}, MP_1^{(8)}, MP_1^{(9)} \right)^T \\ &= \left(FAP_{33}, FAP_{37}, FAP_{20}, FAP_{23}, FAP_{25}, FAP_4, FAP_{52}, FAP_{12}, FAP_6 \right)^T \end{aligned} \tag{4}$$

For other FAPs in a MP, their values could be interpolated by the value of KFAP, as shown in Eq.5. The interpolation coefficient α is a statistic value, which is estimated by Eq.6.

$$MP_j^{(i)} = \alpha_j^{(i)} \bullet KFAP_i \quad (j \ne key_i) \tag{5}$$

$$\alpha_j^{(i)} = \frac{E\left[MP_j^{(i)} \cdot KFAP_i \right]}{E\left[(KFAP_i)^2 \right]} \tag{6}$$

For facial expression synthesis, each KFAP is mapped into a MP by pre-trained values of the interpolation coefficient for FAPs ($\alpha^{(i)}_1$, $\alpha^{(i)}_2$, ... , $\alpha^{(i)}_{Mi}$).

In contrast to the FAPs which directly control the movements of facial points, the MP, which can be represented by KFAP, describes the motion patterns of specific facial organs. It reduces the complexity of FAP manipulation by utilizing the correlation between FAPs, and captures the common expressive facial movements. By using the KFAP, we can model various and even personalized facial expressions, which is useful for both facial expression synthesis and recognition. The different values of KFAP correspond to the continuous change of motion patterns. Fig.2 illustrates the motion pattern of mouth bent.

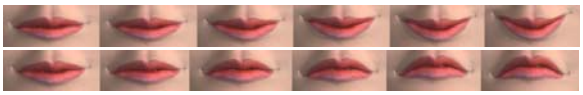


Fig.2. Motion patterns of be bending of lip corners, which could be represented by the value of $KFAP_8$. FAP_{12} is selected as $KFAP_8$, as shown in Table 3.

IV. Modeling the Correlation between Modality Semantics and Facial Expressions

A. Correlation Analysis of SP and KFAP

In order to reveal the correlation between **SP** and **KFAP**, we choose the public Cohn-Kanade facial expression dataset[1] [17].

[1] http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html.

486 facial expression images are randomly selected from Cohn-Kanade dataset as our analysis samples. The selected samples are manually annotated with **SP** and **KFAP** values. We first use the Pearson's correlation coefficient to investigate the correlation within dimensions of **SP** and **KFAP** respectively. First, correlation matrices between the dimensions of SP and between the dimensions of KFAP are calculated based on the above annotated samples. The correlation matrix of dimensions of **SP** is shown in Fig.3(a). The average correlation coefficient between different dimensions is 0.364. The correlation matrix of dimensions of **KFAP** is shown in Fig.3(b). The average correlation coefficient between different dimensions is 0.156. From these results, we can figure out that there is low correlation between dimensions within **SP** and **KFAP**, indicating that there is little overlap or confusion occurs in their own dimensions. It also indicates that **KFAP** has weaker correlation than **SP** (Seen in Fig.3). The reason is that **KFAP** describe the motion patterns of different facial organs which can move independently, while the semantic parameters **SP** delivered by facial expressions are always complex and mixed.

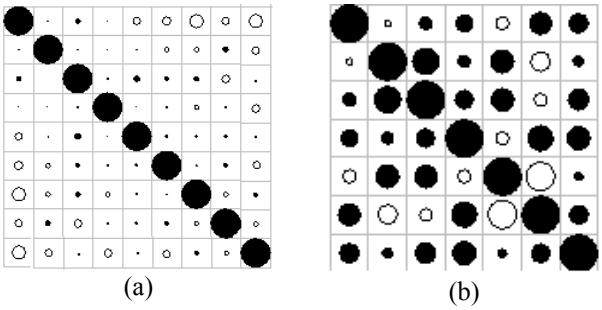


Fig.3. (a) SP correlation matrix. (b) KFAP correlation matrix. (circle area = correlation value, hollow circle indicates negative correlation).

Fig.4 shows the correlation of motion patterns with each of the semantic dimensions. To further investigate the correlation between semantic parameters and motion patterns, the Canonical Correlation Analysis (CCA) is adopted to describe the correlation between **SP** and **KFAP**. CCA is a method to measure the interrelationship between data with multiple variables. The basic idea is to find a space that both dataset can be mapped into, in which their correlation coefficient can be maximized. It is effective in discovering the underlying relationship between two sets of variables. The CCA coefficient calculated this way between **SP** and **KFAP** is 0.864. This result indicates the **SP** and **KFAP** are strongly correlated.

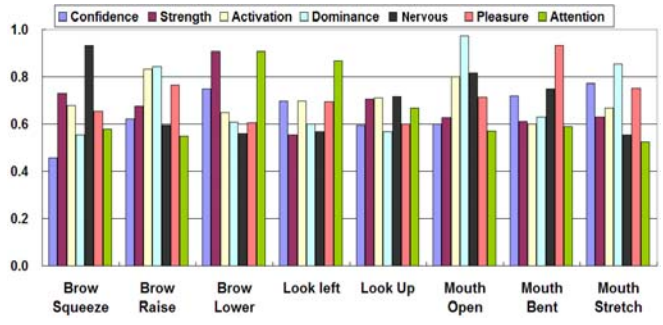


Fig.4. Correlation between semantic parameters and motion patterns.

The above analysis leads to two conclusions: 1) the definitions for both **SP** and **KFAP** are reasonable, due to the low correlation between their own dimensions; 2) there exists strong correlation between modality semantics and facial expressions, indicating the feasibility to model the relationship between **SP** and **KFAP**. Based on these observations, we propose a novel parametric mapping model between modality semantics and facial expressions. The main idea is that, with **SP** as high-level semantic descriptor, **KFAP** applied to low-level expression animation, we build the **SP–KFAP** mapping model.

B. Normalization of KFAP

Actual ranges of motion patterns for different persons are different, but the relative ranges of motion patterns are similar. For example, the KFAP for bend mouth ($MP^{(8)}$) is maximized when laughing, but the values of $KFAP_8$ for different persons laughing are quite different. So we normalize the value of each dimension of **KFAP** to [-1, +1] by Eq.7 and Eq.8, and the normalized version of **KFAP** is used in the mapping models.

$$Normalized(KFAP_i) = \begin{cases} Original(KFAP_i)/F_i^+ & (KFAP_i \geq 0) \\ Original(KFAP)_i/F_i^- & (KFAP_i < 0) \end{cases} \quad where \quad i = 1,...,9 \tag{7}$$

$$F_i^+ = \max(KFAP_i), \quad F_i^- = -\min(KFAP_i) \tag{8}$$

where F_i^+, F_i^- are statistic/empirical values, which define the biggest ranges of $KFAP_i$ in the positive and negative direction respectively. In our work, the values of F_i^+, F_i^- are trained in the Cohn-Kanade dataset. Original values of KFAP can also be retrieved from normalized values with the same F_i^+ and F_i^- using a reverse process.

C. Learning the Mapping between SP and KFAP

As the linear correlation coefficient between SP and KFAP is very strong in CCA analysis, a linear model is first used to simulate the relation between SP and KAFP. Then, a more complex non-linear model is tested. For the linear model, the affine minimum mean square error estimator (AMMSE) is selected, and for the non-linear model, the artificial neural network (ANN) is used.

1. Linear Model

The AMMSE is adopted to obtain the linear mapping from **SP** to **KFAP**, as shown in Eq.9.

$$\mathbf{KFAP} = \mathbf{K}_{ps}\mathbf{K}_s^{-1}(\mathbf{SP} - \mathbf{U}_s) + \mathbf{U}_p \tag{9}$$

where the $\mathbf{K}_s$ and $\mathbf{U}_s$ are the covariance matrix and mean vector of **SP**. The $\mathbf{K}_{ps}$ is the cross-covariance matrix between **KFAP** and **SP**, and $\mathbf{U}_p$ is the mean vector of **KFAP**.

On the other hand, the mapping from **KFAP** to **SP** can be constructed using a similar process:

$$\mathbf{SP} = \mathbf{K}_{ps}^T\mathbf{K}_p^{-1}(\mathbf{KFAP} - \mathbf{U}_p) + \mathbf{U}_s \tag{10}$$

where $\mathbf{K}_p$ is the covariance matrix of **KFAP**, and $\mathbf{K}_{ps}^T$ is the cross-covariance matrix between **SP** and **KFAP**, which is the transpose of $\mathbf{K}_{ps}$.

The advantages of linear mapping are the flexibility and low-complexity in model design and training. Also it can avoid the over-fitting phenomena. However, the correlation between **SP** and **KFAP** may be more complicated than linear relation. So the non-linear model is further considered.

2. Non-linear Model

ANN is adopted to train the non-linear mapping between **SP** and **KFAP**. The feed-forward network composed of one non-linear hidden layer and one linear output layer is designed and illustrated in Fig.5. The neurons in the hidden layer use the transig(•) function as the transfer function, and the linear function is used for the output layer. The number of neurons in the hidden layer is experimentally determined.

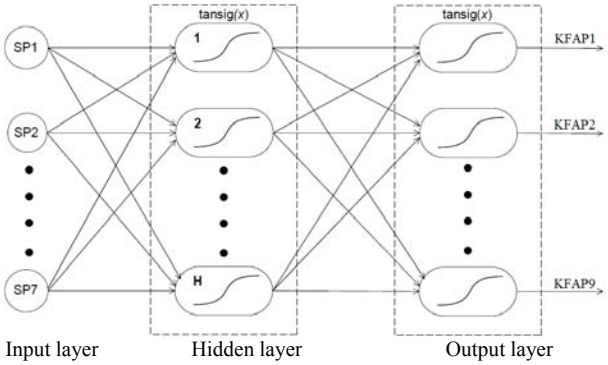


Fig.5. Mapping model between SP and KFAP.

When the input layer is **SP** and output layer is **KFAP**, the number of neurons in the hidden layer is H, the non-linear mapping from **SP** to **KFAP** can be formulated by Eq.11 - Eq.15. $\mathbf{W}^{\text{In}}$ and $\mathbf{b}^{\text{In}}$ are the weight matrix and bias vector on the input layer, and the $\mathbf{W}^{\text{Out}}$ and $\mathbf{b}^{\text{Out}}$ are the weight and bias factor on the output layer. The Levenberg-Marquardt Optimization algorithm is utilized to minimize the mean square error between the estimation and real values of **KFAP**.

$$\mathbf{KFAP} = \mathbf{W}^{\mathbf{Out}}[\text{tansig}(\mathbf{W}^{\mathbf{In}}\mathbf{SP} + \mathbf{b}^{\mathbf{In}}) + \mathbf{b}^{\mathbf{Out}}] \tag{11}$$

$$\mathbf{W}^{\text{In}} = \begin{bmatrix} w_{11}^{\text{In}} & \cdots & w_{1k_s}^{\text{In}} \\ \vdots & \ddots & \vdots \\ w_{H1}^{\text{In}} & \cdots & w_{Hk_s}^{\text{In}} \end{bmatrix} \tag{12}$$

$$\mathbf{b}^{\text{In}} = [b_1^{\text{In}}, b_2^{\text{In}}, \ldots, b_H^{In}]^T \tag{13}$$

$$\mathbf{W}^{\text{Out}} = \begin{bmatrix} w_{11}^{\text{Out}} & \cdots & w_{k_p1}^{\text{Out}} \\ \vdots & \ddots & \vdots \\ w_{1H}^{\text{Out}} & \cdots & w_{k_pH}^{\text{Out}} \end{bmatrix} \tag{14}$$

$$\mathbf{b}^{\text{Out}} = [b_1^{\text{Out}}, b_2^{\text{Out}}, \ldots, b_{k_p}^{\text{Out}}]^T \tag{15}$$

where k_s represents the number of **SP** dimensions, which is 7 in our model. k_p is the number of KFAP dimensions, which is 9 in our model. The number of neurons in the hidden layer H is experimentally determined.

Similarly, when the input layer is **KFAP** and output layer is **SP**, the mapping can be described as:

$$\mathbf{SP} = \mathbf{W}^{\mathbf{Out'}}[\text{tansig}(\mathbf{W}^{\mathbf{In'}}\mathbf{KFAP} + \mathbf{b}^{\mathbf{In'}}) + \mathbf{b}^{\mathbf{Out'}}] \tag{16}$$

where $\mathbf{W}^{\mathbf{In'}}$ and $\mathbf{b}^{\mathbf{In'}}$ are the weight matrix (H' by k_p) and bias vector (H' by 1) on the input layer, and the $\mathbf{W}^{\mathbf{Out'}}$ and $\mathbf{b}^{\mathbf{Out'}}$ are the weight matrix (k_s by H') and bias factor (k_s by 1) on the output layer. H' is number of neutron in the hidden layer, which is also experimentally determined, and possibly different from H.

V. Model Applications

In order to demonstrate the effectiveness of our proposed model, we apply the mapping model to two different applications.

A. Application 1: Facial Expression Synthesis based on Modality Semantics

The **SP** to **KFAP** mapping model is applied in this application. The facial expression synthesis can be integrated into the text to visual speech synthesis (TTVS) system [4][12], which makes the synthetic facial expressions match the semantics implicated by the input text. For each input text, the **SP** is first annotated based on its meaning and the context. Then the annotated **SP** is taken as the input to the **SP** to **KFAP** mapping model, and the estimated **KFAP** is obtained. To animate the facial expressions, we need to calculate all the FAPs in Table 3 depending on the estimated value of **KFAP** according to Eq.5 and Eq.6.

B. Application 2: Modality Semantics Detection from Facial Images

For a new input facial image, FAPs are first detected automatically using a face alignment algorithm [18]. The **KFAP** is constructed from FAP values directly according to Eq.4. Finally, the **SP** is estimated by **KFAP** to **SP** mapping model. This approach enables us to obtain the modality semantics from facial images, which has many potential applications, such as facial expression understanding, semantic-based facial image retrieval and automatic video surveillance.

VI. Experiments and Discussions

In this section, we first verify the reasonability of the semantic parameters (SP) for parameterization of modality semantics. Then we evaluate the effectiveness of the proposed mapping models and their applications by both objective and subjective experiments.

A. Validation on the Parameterization of Modality Semantics

The proposed modality dimensions are validated on a Chinese modality lexical corpus, in which the words can be grouped into synonym groups. An automatic clustering algorithm is applied on the corpus using SP as features, and the results are compared with original synonym groups. The consistency between the clustering results and synonym groups would verify the rationality of the dimensions.

We select commonly used psychological adjectives, mental verbs and modality adverbs [19][20] to form the corpus. 692

Table 4. Confusion matrix between clusters and synonym groups

Clusters	Coherence Percentages $Percent_{i,j}$ (%)									Corresponding synonym Groups
	Sad	***Tolerant***	***Alert***	***Angry***	***Disgust***	***Surprise***	***Fear***	***Believe***	***Happy***	
	G1	***G2***	***G3***	***G4***	***G5***	***G6***	***G7***	***G8***	***G9***	
C 1	**100.0**	—	—	—	—	—	—	—	—	*G* 1
C 2	—	**59.1**	—	4.5	—	13.6	22.7	—	—	*G* 2
C 3	—	—	**64.3**	—	—	28.6	—	—	7.1	*G* 3
C 4	—	—	—	**96.2**	3.8	—	—	—	—	*G* 4
C 5	10.7	—	—	7.1	**78.6**	—	—	—	3.6	*G* 5
C 6	—	—	23.5	—	—	**70.6**	—	—	5.9	*G* 6
C 7	19.2	34.6	—	3.8	—	—	**42.3**	—	—	*G* 7
C 8	—	—	—	—	—	—	—	**84.2**	15.8	*G* 8
C 9	—	—	—	—	—	—	—	—	**100.0**	*G* 9

words are selected in total. For each word, we collect 1) the explanation of its basic sememe which has a modality related meaning; 2) the sample sentence using this word. For example, a modality word in corpus "Romantic", its explanation is "Poetic and full of fantasy", and its sample sentence is "We create a romantic atmosphere". According to [20], these words are manually divided into 9 synonym groups, while each synonym group contains words with similar meanings. Then each word in the Chinese modality lexical corpus is annotated with SP by 5 annotators, according to the explanation of its sememe and sample sentence. Annotation is done with the 5-point Likert scale shown in Table 2. Each annotator is provided with a random subset of the corpus. For each modality word, we ensure it is annotated 2 times at least. It should be noted that all the annotators are researchers who are familiar with Chinese modality semantics, and well-trained in perceptual annotations. For each modality word, if there is an obvious difference among the annotators, they will discuss the case and re-annotate it. The mean score of the annotations are taken as the SP for a modality word. In this way we can obtain SP annotations which are consistent with common human perception.

For validation on the reasonability of the seven modality dimensions, we first adopt the *K*-means clustering algorithm to get the clusters of modality words in our corpus, using SP as clustering features. Then the cluster results are compared with the original synonym groups. Since we have 9 synonym groups in corpus, the number of clusters is preset to be 9. The consistence of the clustering results and the original synonym groups would reflect the rationality of the dimensions.

The 9 cluster results are labeled as C_1, C_2,…, C_9, while the synonym groups are labeled as G_1, G_2, …, G_9. The confusion matrix of the clustering results and synonym groups are illustrated in Table 4, which shows that most clusters nearly correspond to one synonym group. The coherence percentage $Percent_{i,j}$ of cluster C_i and synonym group G_j is defined as Eq.17.

$$Percent_{i,j} = \frac{|C_i \cap G_j|}{|C_i|} \times 100\%, \quad i, j \in \{1,...,9\} \tag{17}$$

where $|\bullet|$ means the number of elements in the set.

These results indicate that the clusters are mostly coherent with the traditional classification of word semantics (i.e. the synonym groups). Therefore, seven modality dimensions defined by us are validated as rational in describing modality information. The result also reveals the effectiveness of the dimensions in describing and distinguishing the modality of words.

B. Dataset and Annotation

Facial Image Dataset: We conduct experiments on the public Cohn-Kanade facial expression dataset. The dataset contains a series of facial expression images of 97 college students, displaying 23 kinds of facial movements. Each image sequence corresponds to a continuous facial movement from neutral state to an extreme state of a kind of facial expression, as shown in Fig.6(a). We randomly select facial expression images with the extreme state from Cohn-Kanade dataset. In total, 486 facial expression images are selected as our experimental dataset.

FAP Annotation: The manual annotation of Facial Fiducal Points on the Cohn-Kanade datasets provided by LAIV laboratory are adopted [21], as shown in Fig.6(b).

(a) (b)

Fig.6. (a) Facial expression images. (b) Facial point annotation.

SP Annotation: Five annotators (2 females and 3 males) are invited to watch the images of facial expressions. Then annotators are asked to rate each facial expression on **SP** using the 5-point Likert scale shown in Table 2. It should be noted that all the annotators are researchers who are familiar with facial expression processing, and well-trained in perceptual annotations. If there is a large difference among the annotators, they will discuss the case and re-annotate it. In this way we can obtain annotations which are consistent with common human perception. The standard deviation of the 5 annotations on an image ranges

from 0.20 to 0.34 on different semantic dimensions. The Pearson correlation coefficients between each pair of annotators' annotations show that there is no significant differences between the annotators ($r=0.46$, $p<0.001$), and thus the mean score of the 5 annotations are taken as the **SP** of each facial expression image.

C. Experimental Setup

Three experiments have been conducted to evaluate the effectiveness of the proposed method and its applications. The first two experiments are designed for facial expression synthesis (Application 1). The third experiment is designed for modality semantic detection from facial images (Application 2).

Experiment 1 is an objective experiment designed for evaluating the **SP** to **KFAP** mapping model and its application to facial expression synthesis. The annotated **SP** are used as input. And the output **KFAP** and the interpolated FAPs are compared with the annotated FAPs to calculate the predicting accuracy.

Experiment 2 is a subjective experiment designed for the application of text-driven facial expression synthesis. We used 80 pieces of text. Each one presents a specific situation. For each piece of text, the **SP** is annotated based on its meaning and the context, by the five annotators introduced in subsection 6.2.1 who also annotated the **SP** for facial expression. The **KFAP** is estimated according to the input **SP**. Then FAPs are interpolated and used to animate the 3D talking avatar to perform the corresponding facial expressions. The 3D talking avatar is introduced by our previous work [4][12].

In order to further prove the effectiveness and necessity of using **SP** for facial expression synthesis, we compare the proposed method with traditional facial expression synthesis method based on six emotion categories (Fear, Surprise, Sad, Angry, Disgust and Happy). The 80 pieces of text using in this experiment are also labeled with one of the six emotion categories by the five annotators. Then FAPs corresponding to the emotion labels are used to synthesize facial expressions.

15 college students (6 females and 9 males) are invited as evaluators in this subjective experiment. The invited evaluators are asked to give Mean Opinion Scores (MOS) according the agreement of the situation described by the texts and the synthetic facial expressions. The synthetic facial expression may be generated based on either modality semantics or emotional labels. And then we compare the scores for our approach and the emotional label approach.

Experiment 3 is an objective test designed for evaluating the **KFAP** to **SP** model which is used in the application of modality semantics detection. The annotated FAPs are used as input. And the output **SP** is compared with the annotated **SP** to calculate the detection accuracy.

For the two objective evaluations, Experiment 1 and Experiment 3, the whole 486 images are randomly divided into 5 subsets with almost the same number of samples in each subset. K-fold cross validation method is adopted using 4 of the subsets for training and one for testing each time. The mean square error (MSE) between estimated values and real values (annotation) are taken as the measurement, as defined in Eq.18 and Eq.19.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - f_i}{\Delta y}\right)^2 \tag{18}$$

$$\Delta y = \max(y_i) - \min(y_i), \quad i = 1,...N \tag{19}$$

where y_i is the real value, f_i is the estimating value, and Δy is the range of real values. Smaller MSE indicates better estimating accuracy. N is the number of samples.

D. Results and Discussions

Experiment 1: We compared the prediction accuracy of **SP** to **KFAP** models trained by AMMSE and ANN respectively. The evaluation results are shown in Table 5. The predicting accuracy of translating **KFAP** to FAPs is also presented, using the annotated values as input.

Table 5. Evaluation results of Application 1

KFAP	MSE		
	SP→KFAP (AMMSE) (x10^{-2})	**SP→KFAP (ANN) (x10^{-2})**	**KFAP→FAP (x10^{-2})**
Raise Eyebrow	2.01	1.90	0.57
Squeeze Eyebrow	3.75	3.38	0.21
Open Eye	3.61	3.39	0.28
Look Left/Right	2.68	2.55	0.19
Look Up/Down	4.28	4.13	0.49
Mouth Open (upper lip)	1.04	0.94	0.08
Mouth Open (bottom lip)	1.95	1.53	0.07
Mouth Bent	1.44	1.23	0.10
Mouth Stretch	2.80	2.49	0.09

The second and third columns compare the predicting accuracy (the average MSE) between AMMSE and ANN. We can find that both AMMSE and ANN can achieve a high predicting accuracy. But ANN obtains a higher predicting accuracy than AMMSE for all KFAPs, especially for Bottom Lip and Mouth Stretch witch are very important to express the degree of modality semantics. The number of hidden units in ANN is experimentally determined as H=7, which obtains the best predicting accuracy. The fourth column of Table 5 is the average MSE of interpolating FAPs by **KFAP**, proving the validity of predicating FAPs by **KFAP** with linear interpolation.

Experiment 2： In this experiment, the **SP** to **KFAP** mapping model trained by AMMSE and ANN respectively are used to map **SP** to **KFAP**, and then translated to FAPs. Based on the MPEG-4 animation framework, the FAPs are used to directly control the 3D facial models to perform the synthetic facial expressions. The (a) and (b) shown in Fig.7 are synthesized by AMMSE and ANN respectively. The synthetic facial expressions driven by the emotion labels (i.e. the six emotion categories) are shown in (c). Sample texts with both emotion label and **SP** are shown in Table 6 and Table 7 respectively.

Table 6. Sample text with emotion label

Sample Text	Emotion label
"You are lying on the bed watching your favorite TV program "	Happy

Table 7. **SP** annotation of sample text

#	Dimension	Value
1	Confidence	0
2	Strength	-0.5
3	Activation	-0.5
4	Dominance	0
5	Nervousness	0
6	Pleasure	1
7	Attention	0.5

15 college students (6 females and 9 males) are invited to take this perceptual evaluation. Each evaluator is required to fully understand the situation described by the texts and then watch and compare the (a), (b), (c) facial expression images. Each facial expression should be scored according to the scale in Table 8 using a MOS method. The average MOS values are 3.53(a), 4.01(b) and 3.36(c) respectively. The one-way ANOVA analysis shows that there exist significant differences among the three groups ($F[2,57]=12.3$, $p<0.005$), and multiple comparison test shows that the B group has a significantly higher MOS than the other two groups. This proves the validity and effectiveness of the proposed **SP** to **KFAP** model trained by non-linear ANN. The results also indicate that the synthetic facial expressions based on **SP** can accurately convey the semantic information from the input texts. As shown in Fig.7, using **SP** as the input, we can characterize more details of facial expressions, and synthesize richer expressions than using emotion categories.

(a) (b) (c)

Fig.7. Samples of synthetic facial expressions:
(a)AMMSE model (b) ANN model (c) Emotion label.

Table 8. The definition of MOS scale

Score	Definition
5	Facial expression fully conveys semantics
4	Facial expression properly conveys semantics
3	Facial expression matches the semantics
2	Facial expression improperly conveys semantics
1	Hard to understand meaning of facial expression

Experiment 3: We build the **KFAP** to **SP** mapping model with both linear and non-linear methods. The predicting accuracy is evaluated by MSE.

The evaluation results are shown in Table 9. The second column illustrates the average MSE using the linear AMMSE model. The third column shows the results obtained by non-linear ANN model. We can find that the ANN obtains much higher predicting accuracy than AMMSE in MSE. The number of hidden units in ANN is experimentally determined as $H=15$, which obtains the best predicting accuracy. In order to demonstrate the effectiveness of our method, we would like to show more synthetic facial expressions on 3D talking avatars in figure 8 and figure 9.

Table 9. MSE of KFAP to SP model tested on experimental dataset

Semantic Dimension	KFAP→SP (AMMSE)	KFAP→SP (ANN)
	MSE ($x10^{-2}$)	MSE ($x10^{-2}$)
Confidence	1.18	1.12
Strength	2.65	2.14
Activation	4.14	2.20
Dominance	2.71	2.68
Nervousness	2.85	1.84
Pleasure	3.19	2.02
Attention	3.24	2.17

VII. Conclusions and Discussions

In this paper, we study the correlation between modality semantics and facial expressions in order to describe and quantify the rich semantics presented by facial expressions. We define and quantify the modality semantics by seven dimensions of semantic parameters (SP), which enable modality semantics to be directly applied to computation. A set of facial parameters called KFAP is used to describe motion patterns of facial organs. The KFAP can be used to generate various and even personalized facial expressions, with lower computing complexity than using MPEG-4 FAP directly. We highlight that there exists a strong correlation between SP and KFAP, indicating the feasibility to model the correlation between them. Based on this analysis, we propose novel mapping models between SP and KFAP, using both linear (AMMSE) and non-linear (ANN) methods. We further apply the proposed models to two different applications: facial expression synthesis based on modality semantics, and modality semantics detection from facial expressions. We conduct experiment to validate the reasonability of the seven modality dimensions. Experimental results on public dataset prove the effectiveness of the models and their applications. The results also indicate that ANN can significantly improve the accuracy in both applications.

The main contributions of this paper are as follows:

1) We formally define the problem of correlation detection between modality semantics and facial expressions. Our statistic analysis verifies the definitions of SP and KFAP, and unveils the strong correlation between them. Empirical studies also show that SP can describe richer expressions than using basic emotion categories, while KFAP can be used to generate natural and vivid facial expressions, with lower computing complexity than using FAP directly.

2) We propose a novel mapping model between SP and KFAP. The model is trained with two different methods: linear regression and artificial neural network (ANN) which is non-linear. Experimental results indicate that non-linear ANN reflects the correlation between SP and KFAP better.

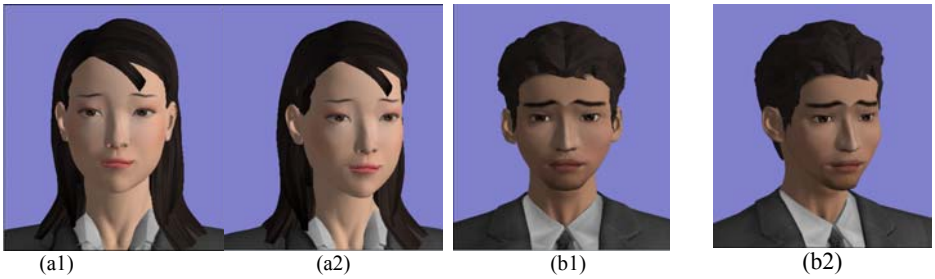
3) To validate the proposed mapping model, we apply it to two applications that motivated our work: facial expression synthesis driven by SP, and modality semantics detection from

facial images. Both objective and subjective experimental results on a public dataset show the accuracy and effectiveness of the proposed model.

Future work may include the scalability of the semantic dimension definition. Another potential issue is to apply the proposed mapping model to other applications (e.g. automatic video surveillance) to further validate its effectiveness.

(a1) (a2) (b1) (b2)

Fig.8. Synthetic facial expressions on different 3D talking avatars, with input SP (1, -0.5, 0.5, 0, 0, 0.5, 0.5). (a1) female, look from the front; (a2) female, look from the side; (b1) male, look from the front; (b2) male, look from the side.



(a1) (a2) (b1) (b2)

Fig.9. Synthetic facial expressions on different 3D talking avatars, with input SP (0.5, -0.5, -0.5, -0.5, 0.5, -0.5, -0.5). (a1) female, look from the front; (a2) female, look from the side; (b1) male, look from the front; (b2) male, look from the side.

ACKNOWLEDGMENT

This work is supported by the National Basic Research Program (973 Program) of China (2011CB302201), National Natural Science Foundation of China (90820304, 60928005, 60931160443). We thank Professor Haizhou Ai from Tsinghua University for providing us the Face Alignment Tools [18], which we use in Application 2 (subsection V.B).

REFERENCES

[1]. Mehrabian, Silent Messages: Implicit Communication of Emotions and Attitudes, 2nd ed. Wadsworth Pub Co.

[2]. P.Ekman, W.V. Friesen, P.Ellsworth, Emotion in the Human Face. New York: Pergamon Press.

[3]. N.H.Frijda, P.Kuipers, E.Terschure, Relations between Emotion, Appraisal, and Emotional Action Readiness, Journal of Personality and Social Psychology, 1989, 57 (2): 212-228.

[4]. Z.Wu, S.Zhang, L. Cai, H.Meng, Real-time Synthesis of Chinese Visual Speech and Facial Expressions using MPEC-4 FAP Features in a Three-Dimensional Avatar, 2006, In: INTERSPEECH(4) :1802 – 1805.

[5]. P. Portner, Modality, Oxford Surveys in Semantics and Pragmatics, 2009.

[6]. Russell, J.A., Fernández Dols,J.M. 1997. The psychology of facial expression. New York: Cambridge University Press.

[7]. Y.Zhang, Q.Ji, Z.Zhu, B.Yi, Dynamic Facial Expression Analysis and Synthesis with MPEG-4 Facial Animation Parameters, IEEE Transactions on Circuits and Systems for Video Technology, 2008,18(10):1383–1396.

[8]. H.Tang, T.S.Huang, MPEG4 Performance-driven Avatar via Robust Facial Motion Tracking, International Conference on Image Processing, 2008, San Diego, CA, USA, 249–252.

[9]. J.Hyun, Y.Lee, Expression Synthesis and Transfer in Parameter Spaces Computer Graphics Forum. 2009, 28(7):1829-1835.

[10]. M.Z.Uddin, An Enhanced Independent Component-Based Human Facial Expression Recognition from Video. IEEE Transactions on Consumer Electronics, 2009,55(4):2216-2224.

[11]. Y.Shin, Y.Kim, E.Y.Kim, Automatic Textile Image Annotation by Predicting Emotional Concepts from Visual Features, Image and Vision Computing, 2010, 28(3):526-537.

[12]. J.Jia, S.Zhang, F.Meng, Y.Wang, L.Cai, Emotional Audio-Visual Speech Synthesis based on PAD IEEE Transaction on Audio, Speech and Language Processing, Digital Object Identifier: 10.1109/TASL.2010.2052246.

[13]. M.Obaid, R.Mukundan, M.Billinghurst, M.Facial, Expression Representation using a Quadratic Deformation Model, Computer Graphics, Imaging and Visualization, 2009.

[14]. Motion Pictures Expert Group: ISO/IEC 14496-2: International Standard, Information Technology-Coding of Audio-visual Objects, part 2: Visual; amendment 1: Visual extensions (1999/Amd. 1: 2000(E)).

[15]. F.Lavagetto, R.Pockaj, An Efficient use of MPEG-4 FAP Interpolation for Facial Animation at 70 bits/frame, IEEE Transactions on Circuits and Systems for Video Technology, 2001,11(10): 1085–1097.

[16]. S.Zhang, J.Jia, X.Wang, L.Cai. Facial Expression Synthesis based on Semantic Dimensions. Qinghua Daxue Xuebao, 2011, 51(1): 80-84. (In Chinese).

[17]. T.Kanade, J.Cohn, Y.Tian, Comprehensive Database for Facial Expression Analysis, In Proc 4th IEEE International Conference on Automatic Face and Gesture Recognition, 2000, IEEE Press, 46-53.

[18]. L.Zhang, H.Ai, *et al.*, Robust Face Alignment Based on Local Texture Classifiers, The IEEE International Conference on Image Processing, 2005, Italy (2):354-357.

[19]. L.Peng, On Modality of Modern Chinese, Fudan University, Shanghai, 2005.

[20]. J.Mei, Synonyms Set, Shanghai Ci Shu Press, Shanghai, 1983.

[21]. G.Lipori, Manual Annotations of Facial Fiducial Points on the Cohn Kanade database, 2009-11-10.

[22]. A.Mehrabian, Framework for a Comprehensive Description and Measurement of Emotional States, Genet Soc Gen Psycol Monogr, 1995, 121(3): 339-361.