Facial Expression Recognition and Synthesis toward Construction of Quasi-Personality

JUN HAKURA, MAMORU KASHIWAKURA,
YOUICH HIYAMA, MASAKI KUREMATSU, and HAMIDO FUJITA
Faculty of Software and Information Science
Iwate Prefectural University
152-52 Sugo, Takizawa-mura, Iwate
JAPAN
http://www.fujita.soft.iwate-pu.ac.jp/en/index.html

Abstract: - This paper describes the visual components of the Virtual Miyazawa Kenji Project [1] that aims to restore Japanese famous writer, poet, and essayist Miyazawa Kenji as an artificial system. The method introduced can be adopted to the other person who have left behind works with deep insights. Namely, a construction of quasi-personality in computers is the main target of the project. For this aim, this paper tries to achieve a facial expression recognition method and synthesis method by means of the same discipline. This discipline is implemented as a database named Facial Expression Database (FED). A linear system identification approach [3] is introduced to represent the typical movements observed in the facial expressions. The concrete methods and preliminary implementations are shown in the paper.

Key-Words: - Facial Expression Database, Facial Expression Recognition, Facial Expression Synthesis, Miyazawa Kenji Project, Avatar, Linear System Identification

1 Introduction

The requirements for mechanisms that enable easy interactions between human and artifacts to make the human easy to access the services the artifacts provide, are extremely high in the today's ubiquitous computerized world. One of these mechanisms that are definitive to the artifacts is reading the mental state of the human in natural conversations. One of the key features is said to be in the human facial expressions. We can guess the particular mental state of that person without any word, but from his/her face. The same function would make the artifacts to be able to understand the human more accurate.

Another mechanism to be required is synthesizing the quasi-mental states of the artifacts through the same channel, i.e., facial expressions. This would enhance the human requesting or accessing the artifact, especially for those who are not accustomed to such artifacts.

This paper tries to achieve the both mechanisms with the same discipline. This means that the system uses the same rules to recognize and synthesize facial expressions. The face is considered as the three systems with basically six modes (each mode might consist of several sub-modes) that correspond to the six emotions, i.e., joy, sadness, anger, disgust, fear, and surprise [2]. Therefore, a system identification approach to facial expression recognition [3] is introduced in the paper.

2 Facial Expression Recognition and Synthesis for a Quasi-Personality

The main target of the paper is construction of visual components for Virtual Miyazawa Kenji System [1]. The system tries to reconstruct the cognitive aspects of Japanese famous writer, poet, and essayist Kenji Miyazawa (Kenji is his first name, and Miyazawa is his last name) who left behind his deep insights in his works. This paper regards the cognitive thinking style of a person as ones personality. Therefore, the artifact that realizes the personality is considered as quasi-personality.

Although there are effective studies for each of facial expression recognition and synthesis mechanisms, in most studies, the mechanisms are separated for the inconsistency among the human recognition [4, 5], and synthesis [6] of the facial expressions. This means that each mechanism is optimized for the functional aspects of its aim so that reliable and acceptable recognition/synthesis are achieved.

However, constructing a quasi-personality that tries to computationally and virtually reproduce a highly spiritual person who has a plenty of works on his thought, requires the both aspects altogether. A person is considered as to recognize the facial expressions according to his/her mental images of his/her own facial expressions. Namely, we assume that a person recognizes and exhibits the facial expressions based on the same rules. Recently, under benefits of Active Appearance Model (AAM) [7], the analysis and synthesis is done by means of the same discipline [8]. However, the synthesized faces are analogous to the original face but could be recognized as a different person.

Thus, this study tries to construct a database that relates the movements of the facial feature point with the six emotions that Ekman advocates [2]. A linear system identification approach is introduced to extract the corresponding movements for each of the six emotions as described in Section 3. The movements are stored in the form of transition matrixes that estimates the locations of the facial feature points in the next time step using the locations in the current time step. The labeling of the facial expressions is done manually by means of FACS (Facial Action Coding System). The database is used in the both recognition and exhibition process. Therefore, the both processes are conducted by the same discipline.

3 Facial Expression Recognition and Synthesis Based on FED

The recognition and the synthesis are the reciprocal processes in this paper. They are both relied on the Facial Expression Database (FED). To achieve the processes, mainly three methodologies are introduced, i.e., a construction method of FED, a facial expression recognition method, and a facial expression synthesis method based on the FED, as shown in Fig.1. This section tries to describe each method in detail.

3.1 Facial Expression Database (FED)

Facial expressions in this study are considered as results of the movements of facial feature points. Facial feature points are the points that are enough to estimate the target emotional expressions by observing the movements of the points. The movements of the feature points are assumed to be caused by a set of systems as in the Facial Score [3]. By identifying the systems for six basic categories [2], we can obtain the knowledge on the movements of the facial points to

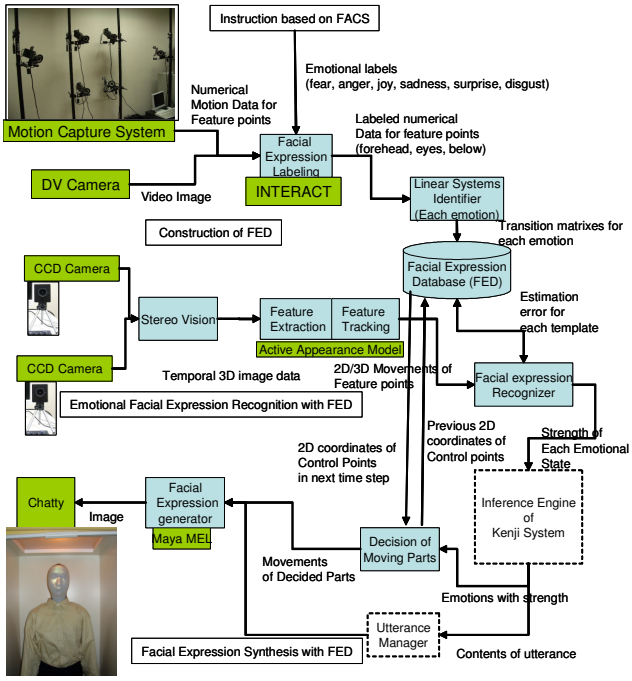

Fig. 1 System Block Diagram of
the Visual Components of Kenji System
(Dashed Boxes are out of the range of this paper.)

synthesize Kenji's emotional facial expressions and to recognize the user's emotions. The system identification method, i.e., LSM: Least-Square Method, is adopted to identify the systems. A face is divided into roughly three parts in this paper: eye brows, eyes, and mouth. This will allow the Virtual Kenji system to recognize mixed emotions as FACS recognizes the mixed emotions with combinations of 44 Action Units. Each of these parts is assumed to contain its own emotional signals. The results of the identification for each part would be the six transition matrixes; each of those can estimate the movements of the facial feature points for particular emotional facial expressions. The acquired knowledge is collected as a database, called Facial Expression Database (FED).

To extract typical movements of the facial feature points that express the particular emotion, we should collect facial expressions that represent the particular emotion. For this aim, subjects who have trained to express the six emotions act each emotion for several times. A motion capture system and a digital video camera are used for the observation. The motion capture system extract precise trajectories of the marker attached at the feature points. The digital video camera captures the facial expressions of the subjects at the same time. A third person watches the video and label the durations where the particular emotions are

expressed. The person relies on the FACS to label the durations so that the label is used for general users.

The typical movements of the points are then to be extracted from the durations labeled as the same emotion. As mentioned above, the typical movements are represented as a result of a system's output. Then, the face in this paper consists of three systems: eye brows, eyes, and mouth. The system has some modes; initially, every duration is assumed as a mode. A merging algorithm of the modes reduces the number of the modes. The similar movements are considered as results of the system in the same mode. Namely, each mode expresses an emotional category. The algorithm is described in the previous paper [1].

As the results of applying the algorithm, the each mode corresponds to the most typical movements of the points to express an emotion. For example, when there exists N durations labeled as "fear", the modes that representing fear are described as follows:

$$\mathbf{M}^{\mathbf{fear}} = \{\mathrm{M}^{\mathbf{brows}}, \mathrm{M}^{\mathbf{eyes}}, \mathrm{M}^{\mathbf{mouth}}\},$$
$$\mathrm{M}^{p} = \{M_i^p \mid 1 \le i \le m, i \in I, 1 \le m \le N\}. \quad (1)$$

where, $p = \{brows, eyes, mouth\}$, I denotes a set of positive integers. M^p is obtained for each of the six emotions.

The modes are represented as the transition matrixes. Namely, every facial feature point can be estimated by using the matrixes. Let eye brows requires six feature points (three points for each brow) at time t denoted as $FP^p(t) = (px_1, py_1, px_2, py_2, \cdots, px_6, py_6)^T$, where px_i and py_i are the x and y coordinate of i-th feature point respectively. Then, the points at the next time step can be estimated by the following equation:

$$FP_i^{p*}(t+1) = M_i^p \bullet FP^p(t) \ . \quad (2)$$

where, p* means that it is the estimated value. Note that M_i^p is a 12 x 12 matrix in this example. Note also that every estimation value is calculated by the actual value of the facial points. The recognition process compares the error between actual and estimated values of the facial points for every mode, part, and emotion. The next section describes detailed method to make recognition with the modes.

3.2 Facial Expression Recognition with FED

The emotional expressions to be recognized in this paper are the status of the six emotions at certain time. Therefore, the system introduces a model of the emotional states of a user. This model initializes

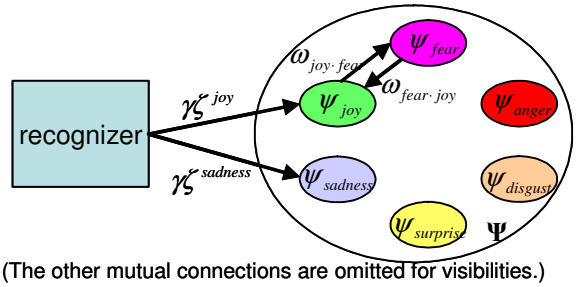

Fig. 2 Mental Estimation Model for User

accumulated user states, every time a user leaves from in front of the Virtual Kenji system. The outline of the model is depicted in Fig. 2. It is implemented as a weighted network with the matching rates described bellow as inputs and the state vectors as outputs. The weights in the network are tuned appropriately from psychological points of view.

In the recognition process, the stereo vision system is introduced to track the facial feature points without any marker on the user's face. Considering the usage of the system will lead to this conclusion, in spite of the computational costs, weak tracking accuracy owe to the brittleness to the environmental noises. Namely, attaching markers on the face harms the comforts and usabilities of the system. Fortunately, the AAM provides detectability of the points from the images/ video sequences. There are several applications of the AAM that track certain objects (e.g. [9]). There are also several applications that apply the AAM to 3D images [10]. These precedence technologies enabled displacement of the motion capture system to the stereo vision system in the recognition process. In the database construction process, we still use the motion capture system for its accuracy in tracking.

As mentioned in the previous section, FED provides system identifiers of the systems that control the facial feature points. Therefore, the facial expression recognition with FED uses these identifiers to know to what emotional categories the presented facial expression belongs. The presented facial expressions can be detected as the movements of the facial feature points with stereo vision system and the AAM, so that $FP^p(t)$ in Equation (2) are available at every time step. Every identifier estimates $FP_i^{p*}(t+1)$ at the next time step through Equation (2). These estimated points are compared with the actually observed points at t+1, then we can get the errors e_i^p :

$$e_i^p = (E_i^p)^T (E_i^p),$$
$$E_i^p = FP_i^{p*}(t+1) - FP^p(t+1). \quad (3)$$

Note that e_i^p is a scalar value, and E_i^p is a vector. Then, according to Equation (1), we have $|M^p|(=m)$ errors for each facial part with respect to each emotional category. To determine which emotion is observed, we have to cumulate the error values of each part. We simply employ the minimum value for the aim, because the identifier with the minimum error itself implies that the emotion is detected:

$$e^p = \min_i \{e_i^p\} \quad (4)$$

Namely, we have now error vectors Δ_e consist of three elements, i.e., on eye brows, eyes, and mouth, for each emotion:

$$\Delta_e = (e^{brows}, e^{eyes}, e^{mouth}),$$
$$e \in \{\text{joy}, \text{sadness}, \text{anger}, \text{disgust}, \text{fear}, \text{surprise}\} \quad (5)$$

The elements of the error vectors considered to be reflecting the degrees of match between the modes and the expressed emotions. Therefore, the matching rate ζ_p is defined as follows:

$$\zeta_p = \frac{1}{1+\alpha(e^p)^2} \quad (6)$$

where, α is a coefficient. The facial parts and their contributions to the emotional expressions might be different in each emotion. Therefore, matching rate ζ^e of certain emotion for the whole face is calculated by weighted sum of each ζ_p:

$$\zeta^e = \sum_p w_p^e \zeta_p \quad (7)$$

where, w_p^e is the weight on emotion e at facial part p, and $\sum_p w_p^e = 1$.

These matching rates for the emotions are the inputs to the model of the emotional state $\boldsymbol{\Psi}$ of the user. $\boldsymbol{\Psi}$ is a vector that has six emotional states values ψ_e as its elements. ψ_e is a time variant and therefore it should be denoted with a time index:

$$\psi_e(t+1) = l(\beta \cdot \psi_e(t) + \gamma \cdot \zeta^e + \mu \cdot \sum_{i \neq e} \omega_{ie} \cdot \psi_i(t)) \quad (8)$$

where,

$$l(x) = \begin{cases} x\,(x \geq 0) \\ 0\,(x < 0) \end{cases},$$

β is a damping coefficient, γ, μ are coefficients, and ω_{ie} is a weight on the connections from state i to e. The emotional state vector is sent to the central inference engine of the Kenji system. The engine determines the emotional output of the whole system according to the Kenji Style [1]. The description on the inference engine is beyond the scope of the paper. The rest of this section is assigned for the synthesis method of the facial expressions of the Kenji system.

3.3 Facial Expression Synthesis with FED

Facial expressions should contain not only a single emotional sign but also mixture of such signs. The FACS nicely utilizes those signs by dividing the facial parts with their movements as AUs (Action Units) [2]. Combinations of the AUs are representing not only a single emotion but also the mixture of those emotions. However, the trajectories of the movements of the parts are ignored in the FACS. The trajectories of the feature points will be useful especially in synthesizing the facial expressions. The FACS relates the AUs with the six emotions. This paper utilizes the relationships to synthesize the facial expression and FED is used to control the movements of the AUs. The AUs in this paper are categorized into two categories CA1 and CA2: CA1 is the category that has strong relations with active emotional expressions, and CA2 is not.

Fig. 3 depicts the outline of the synthesis component. This component receives an emotion vector as an

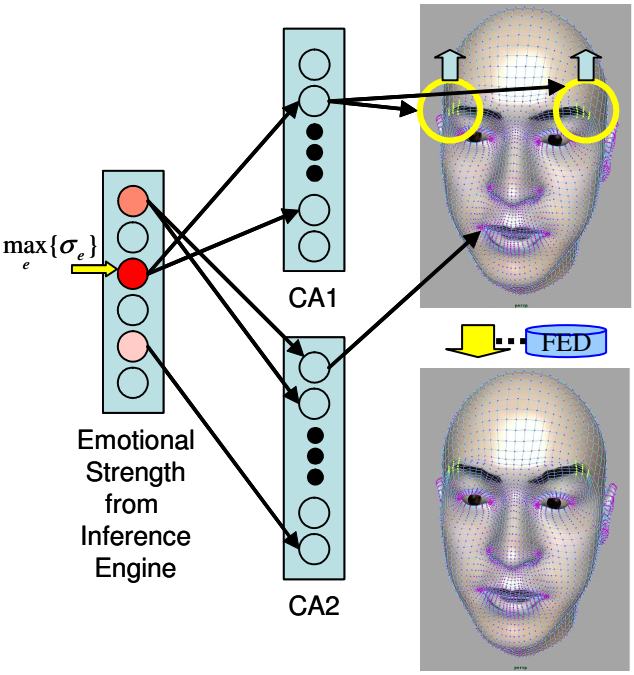

Fig. 3 Outline of Synthesis Component.

input. The emotion vector has the six emotional states as elements that inferred in the inference engine of the Virtual Kenji. Each emotional state has a strength value $\sigma_e (0 \leq \sigma_e \leq 1)$. The emotional states are linked with the corresponding AUs in reference to the FACS. AUs are also linked with the corresponding facial parts of the 3D facial model of the Virtual Kenji. Each facial part has its own control points that are placed around the facial feature points described in the previous subsections. Each control point has a boundary of the movement.

The AUs of CA1 corresponding to the emotional state that has maximum strength is the target of the expression. The other emotions are expressed by means of the AUs of CA2, in proportion to the strengths. The AUs of CA1 make the corresponding parts of the 3D facial model moving. The trajectories are also calculated by means of Equation (2) within the boundaries in proportion to the strength.

In the next section, preliminary implementations of some of the visual components of Virtual Kenji system with brief result examples are described.

4. Preliminary Implementations

Miyazawa Kenji project is an ongoing project and therefore, the system is still under construction. Several implementations, however, are already done and some of which is described with some results in this section. More results will be shown in the conference site.

Fig. 1 is a block diagram concerning with the visual components of system. As shown in the figure, the hardware of the system consists of the motion capture system manufactured by MotionAnalysis Corp., a stereo vision system by Point Grey Research Inc., work stations as controllers, and the Chatty by Ishikawa Optics& Arts Corp., i.e., a lay figure that has a screen on the face and able to display the facial video images on it. In addition to the software components described above, we introduce several software products, e.g., INTERACT by Mangold International GmbH to label the facial expressions, and Maya and its script language Maya MEL by Autodesk Inc., to construct the facial model of Miyazawa Kenji and to control the facial expression of the model, respectively.

Fig. 4 shows an experimental environment for construction of FED. In this preliminary experiment, a subject acts five facial expressions of five emotions, i.e., fear, joy, sadness, surprise, and disgust, from the six emotions. Throughout the experiment, the only the

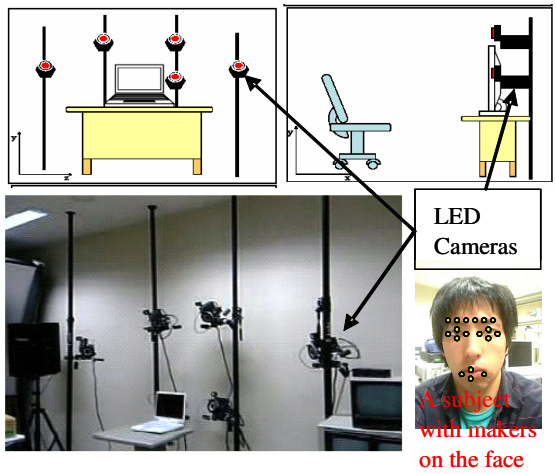

Fig. 4 An Experimental Environment

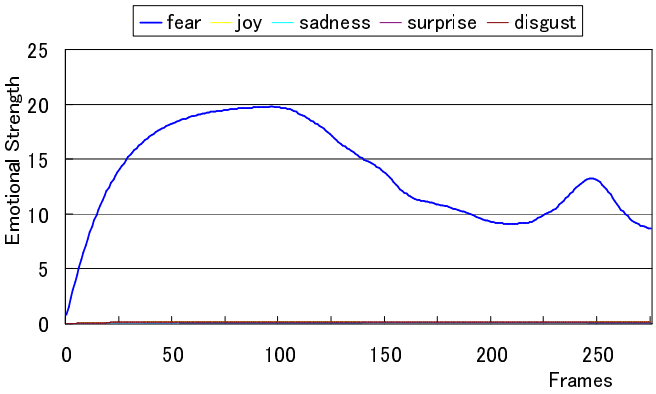

Fig. 5 Transitions of Estimated Emotional State Value for Facial Expression of Fear at Experiment

movements of eye brows are used. After construction of FED based on the proposed method, the expression for fear used in the construction phase is tried to be recognized for the subject. For this experiment, we set the parameters as follows: $w_p^e = 1$ (iff p = "eye brows"), $w_p^e = 0$ (elsewhere), $\alpha = 1$, $\beta = 0.95$, $\mu = 0$. Fig. 5 shows the transition of each value of the estimated emotional state value of the subject. As shown in the graph the value for fear grows greater than the other emotional state values.

Fig. 6 displays the six emotions by Virtual Kenji constructed by means of the proposed method. As shown in the figure, the emotions are nicely displayed on the face. There would be no doubt that each emotion is displayed by the same person.

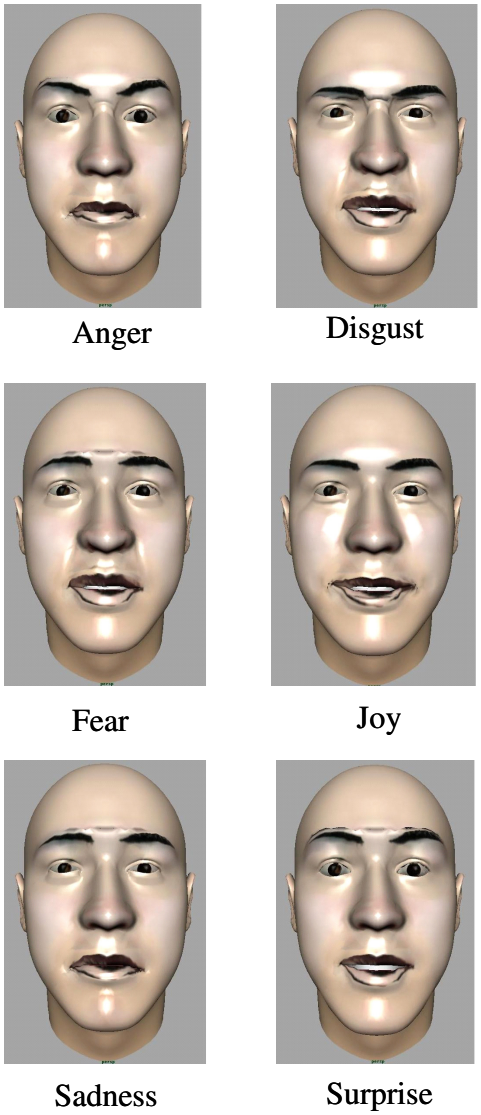

Fig. 6 Synthesized Facial Expressions for the Six Emotions

5. Conclusion

This paper tries to achieve a facial expression recognition method and synthesis method by means of the same discipline. This discipline is implemented as a database named Facial Expression Database (FED). A linear system identification approach is introduced to represent the typical movements observed in the facial expressions. The concrete methods and preliminary implementations are described in the paper.

Descriptions on the preliminary implementation indicate possibilities of the proposed methods.

Acknowledgement:
This work is supported by a grant from Research and Regional Cooperation Division, Iwate Prefectural University, with which Hamido Fujita is the principal investigator.

.

References:

[1] H. Fujita, J. Hakura, M. Kurematsu, Virtual Cognitive Model for Miyazawa Kenji Based on Speech and Facial Images Recognition, *WSEAS Transactions on Circuits and Systems*, Issue 10, Vol.5, 2006, pp. 1536-1543.

[2] P. Ekman, and W. V. Friesen, *Unmasking the Face*, Prentice Hall, NY, 1975.

[3] M Nishiyama, et. al., Facial Expression Representation Based on Timing Structures in Faces, *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2005, pp. 140-154.

[4] M. Pantic and L J. M. Rothkrantz, Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, No. 12, 2000, pp. 1424-1445.

[5] J. R. Movellan and M. S. Bartlett. The next generation of automatic facial expression measurement. *In Paul Ekman, editor, What the Face Reveals*. Oxford University Press, 2003.

[6] Q. Zhang, et. al., Geometry-Driven Photorealistic Facial Expression Synthesis, *Eurographics/ SIGGRAPH Symposium on Computer Animation*, 2003.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor, Active Appearance Models, *Proc. of European Conference on Computer Vision*, Vol. 2, 1998, pp. 484-498.

[8] B. Abboud, F. Davoine, M. Dang, Expressive Face Recognition and Synthesis, *IEEE workshop on Computer Vision and Pattern Recognition for Human Computer Interaction*, 2003.

[9] D. Witzner Hansen and A. E. C. Pece, Eye Tracking in the Wild, *Computer Vision and Image Understanding*, Vol. 98, No. 1, 2005, pp. 155-181.

[10] R. Gross, I. Matthews, S. Baker, Active Appearance Models with Occlusion, Image and Vision Computing, Vol. 24, No. 6, 2006, pp. 593-604.