An Exploration toward Automatic Facial Expression Analysis for Intelligent HCI Systems

JUN HAKURA, MASAKI KUREMATSU, HAMIDO FUJITA
Faculty of Software and Information Science
Iwate Prefectural University
152-52 Sugo, Takizawa, Iwate 020-0193
JAPAN

Abstract: - This paper describes our ongoing exploration toward facial expression analysis for intelligent HCI systems. Two approaches are proposed from two different points of views. The first approach captures dynamic movement accompanied with facial expressions as the clue for estimating emotion. In this approach, the movements are considered as the results of activities by virtual systems that are assumed to be embedded in the face. We call this approach as System Identification Approach. The second approach relies on the positions of the facial parts, such as eyes, eye brows, and mouth, to detect emotion. This approach, named Facial Expression Map Approach, relates the positions of the parts with the six basic emotions and the relations are represented as maps. The outlines of the methods based on the approaches are described with some experimental results.

Key-Words: - Facial Expression Analysis, Emotion Estimation, Emotion Recognition, Facial Expression Map, System Identification, HCI, Intelligent Interface.

1 Introduction

This paper introduces our ongoing exploration on automatic facial expressions analysis of unspecified persons for the sake of building intelligent HCI (Human Computer Interaction) systems. As an example of intelligent HCI systems, Miyazawa Kenji (MK, hereafter) Project, e.g., [1], [2], tries to artificially revive Japanese famous writer/poet MK.

Automatic analysis of facial expressions that aims to estimate emotions from facial expression so far has no definitive method, although there are considerable amount of studies with certain results (e.g., [3], [4]). Most of these studies rely on FACS (Facial Action Coding System) [5] that provides a categorical approach to the facial expression analysis. Namely, they relate facial expressions with categories, i.e., the basic six emotions, by observing AUs (Action Units).

This study tries to realize the automatic facial expression analysis also within the categorical approach. As discussed in the next section, however, the following additional three characteristics are quested to enhance applicability to HCI: (1) estimative ability of emotional strengths, (2) detectabitily of the starting point of the emotion (facial display), and (3) irrelevance to the facial expressions of the user at the beginning of the interaction. To achieve the three characteristics two approaches are introduced: System Identification approach and Facial Expression Map approach. In the rest of the paper, the outlines of the methods based on the approaches are described with experimental results.

2 Emotion Estimation from Face

This section briefly discusses importance of the three characteristics mentioned in the previous section, for Intelligent HCI system. Here we call systems with emotional interpretation abilities and responsive to the emotional information as Intelligent HCI systems.

2.1 Estimative Ability of Emotional Strength

The intelligent HCI systems in the above sense requires sensitive response to the subtle changes in the facial expressions of the user. The basic six emotions have several meanings with respect to the their strengths. Thus, the systems are to be responsive to these strengths. We assume here that the magnitudes of the expressions have relations with the emotional strengths. This characteristic is mainly realized by Facial Feature Map Approach described in 3.3.

2.2 Detectability of Starting Point

For the systems like we are aiming here, the exact timing of when the user react to the system is important. What makes the HCI systems intelligent

might be an ability to reason preferences of the user through the interaction, and act in consideration of them at the similar occasions. The starting point of the emotional facial expression would help the reasoning. Thus the detectability of the starting point of the emotional expressions is considered as one of the important characteristics. This characteristic is mainly achieved by System Identification Approach in 3.2.

2.3 Irrelevance to First Glance Expressions

This characteristic is more technical problem than the previous two. Systems like MK system should interact with general public. This implies not only the system should do with a million of faces, but the user is not guaranteed to stand in front of the system with the neutral face that apt to be used as the yardstick to detect AUs. The system are required to recognize emotions without the yardsticks. The both approaches proposed here overcomes the problem: System Identification Approach solves it by focusing on the movements, while Facial Expression Map Approach solves it by directly comparing the expressions with the maps as described in the following section.

3 Facial Expression Analysis Methods

The proposed methods can be distinguished according to what is to be the facial displays to be associated with the emotion. We consider there are two: one is the movements of the facial parts such as eyes, eye brows, and mouth, and another is the positions of the same parts. To make their computations easier, like most other methods, the both methods adopt a set of feature points with which positions and movements the relevant facial parts are perceived.

This section first describes the outline of the common architecture. Then, detailed mechanisms of each method are described respectively.

3.1 Common Architecture

The common Architecture for our facial expression analysis method consists of roughly four components as shown in Fig.1, i.e., feature point extraction component, feature point tracking component, facial expression analyzer, and emotional state estimator. The image flows from camera captures facial expressions of the user. Here, we assume that every image contains a face. Then, a set of feature points is to be extracted in the image. Temporal transitions of these feature points contain the information on facial behaviors. Thus these feature points are to be tracked at every moment. The feature points, then, sent to the analyzer to reveal the amounts of the emotions implied by the expression. The estimated amounts are sent to the estimator which infers the emotional states of the user by taking the temporal changes on them into account. The rest of the subsection briefly describes each method of feature extraction, feature tracking, and emotional state estimator.

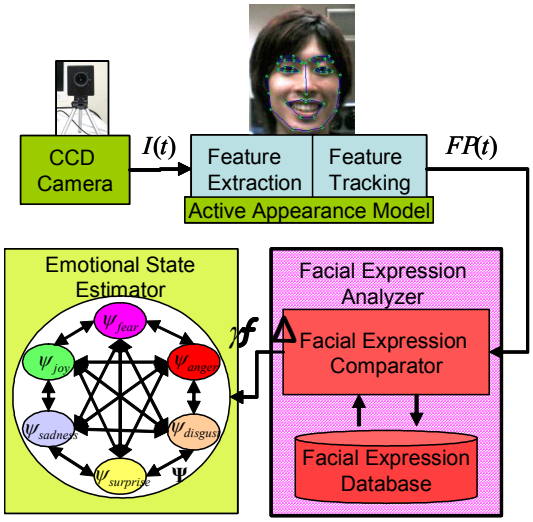


Fig.1 Outline of Common Architecture

3.1.1 Feature Extraction and Tracking Method

The feature extraction and tracking in this paper is done at the same process by means of the AAM (Active Appearance Model)[6]. Although there is a defect in the accuracy, there are several advantages with adopting the AAM, such as free from markers, and availabilities of the device and real time processing. The AAM is a statistical model/method to detect deformable objects in images. It makes a statistical model of the shapes and appearances from the training images with manually added feature points. When applied to novel images, it finds parameters to adapt the model to the novel image by means of a optimization techniques. Today's advance of computational devices enables the method to be processed in the real time so that it can be applied to the object tracking tasks. The feature points on the face in Fig.1 are automatically detected by AAM.

3.1.2 Emotional State Estimation

As described in the next subsection, the role of the facial expression analyzer is detecting to what extent the expression contains the basic six emotions. Thus the detected emotions are temporal and might contain misunderstandings of the expressions, because the analyzer tries to detect emotion from every frame.

A mechanism that could provide more stable estimations of the users' emotional states so that they can be used in the systems interacting with users, is required. For this reason, emotional state estimator is introduced. the estimator is inspired by the mental model known to as Cathexis model [7]. The estimator consists of state nodes corresponding to the basic six emotions, weighted links among them, and also weighted links between the detected emotional values in the analyzer and corresponding state nodes. The state nodes possess the estimated emotional values ψ_e, calculated by means of the overall mechanism.

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

where, $l(x) = x(x \geq 0); 0(x < 0)$, β is a damping coefficient, γ, μ are coefficients, ω_{ie} is a weight on the connections from state i to e, and $\zeta^e(t)$ is the detected emotional amounts at time t in the analyzer, as described in the next subsection.

As mentioned above, two approaches can be embedded as the facial expression analyzers. System Identification Approach relies on the movements of the facial feature points, Facial Feature Map Approach, on the other side, relies on the positions of these points. The rest of this chapter concentrates on the description of the detailed mechanisms and current results achieved by each method.

3.2 System Identification Approach

Facial expressions in the system identification approach are considered as results of the movements of the facial feature points. These movements are assumed to be caused by a set of systems as in the Facial Score [8]. Main difference between the Facial Score and the proposed method is that we relate the modes of the systems directly with the emotions so that they can be used for automatic analysis of facial expressions. The system identification method, i.e., LSM: Least-Square Method, is adopted to identify modes of the systems. Each mode of the identified system associated with one of the six basic emotions by means of the FACS. These transition matrices being coupled with corresponding emotional labels, are stored in a database named as Facial Expression Database (FED). Transition matrices that represent these mode can be used to estimate the positions of the feature points in the next time step. Thus, by comparing the estimated positions with those observed in the tracking process, we can estimate what emotion can be detected for the image frame.

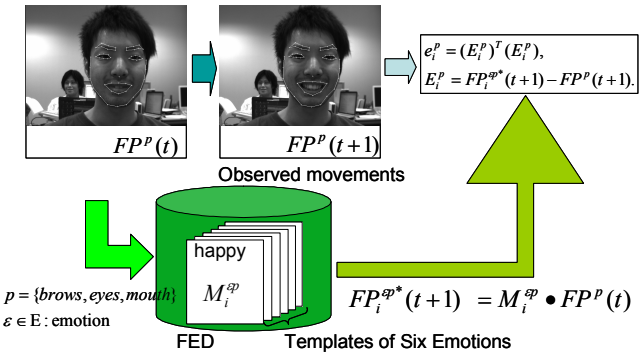


Fig.2 Outline of System Identification Approach

The rest of section describes FED, emotion estimation method with FED, and an experimental result with the system identification approach.

3.2.1 Facial Expression Database for System Identification Approach

A face is divided into roughly three parts in this paper: eye brows, eyes, and mouth. Each part is assumed to contain its own emotional signals. The results of the identification for each part would be the six transition matrices; each of those can estimate the movements of the facial feature points for particular emotional facial expressions. The acquired matrices are 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 expressions that represent particular emotions. For this aim, subjects who have trained to express the basic six emotions, act each emotion for several times, and the movements at that time are labeled as the emotion. As depicted in Fig.3, the movements are represented as a result of a system's output. The system has modes; initially, whole durations of the training frames are assumed to be 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 detail of the algorithm is described in [2]. By applying the algorithm, each mode are corresponded with the most typical movement expressing the emotion. For example, assuming N durations labeled as "fear", the modes that representing fear are described as follows:

$$\mathbf{M}^{\mathbf{fear}} = \{\mathrm{M}^{\mathbf{brows}}, \mathrm{M}^{\mathbf{eyes}}, \mathrm{M}^{\mathbf{mouth}}\},$$
$$\mathrm{M}^p = \{M_i^p \mid 1 \leq i \leq m, i \in I, 1 \leq m \leq N\}. \qquad (2)$$

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

The modes are represented as the transition matrices. Namely, every facial feature point can be estimated by multiplying the matrices with current positions of the feature points. 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:

$$FP_i^{p*}(t+1) = M_i^p \bullet FP^p(t) \quad (3)$$

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 estimation 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 estimation with the modes.

3.2.2 Emotion Estimation with FED

As mentioned in the previous section, FED provides identified systems that control the facial feature points. Therefore, emotional estimation in this method uses these systems to categorize the facial expressions. The presented facial expressions can be detected as the movements of the facial feature points with a vision system and the AAM, so that $FP^p(t)$ in Equation (3) are available at every time step. Every system estimates $FP_i^{p*}(t+1)$ with Equation(3). The estimated points are compared with the actual observed points at t+1, to calculate the error e_i^p:

$$\begin{aligned} e_i^p &= (E_i^p)^T (E_i^p), \\ E_i^p &= FP_i^{p*}(t+1) - FP^p(t+1). \end{aligned} \quad (4)$$

Note that e_i^p is a scalar value, and E_i^p is a vector. According to Equation (2), we have $|M^p| (= m)$ errors for each facial part with respect to each emotion. 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:

$$e^p = \min_i \{e_i^p\} \quad (5)$$

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

$$\begin{aligned} &\Delta_\varepsilon = (e^{brows}, e^{eyes}, e^{mouth}), \\ &\varepsilon \in \{\text{joy, sadness, anger, disgust, fear, surprise}\} \end{aligned} \quad (6)$$

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:

$$\zeta_p = \frac{1}{1 + \alpha (e^p)^2} \quad (7)$$

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

$$\zeta^e = \sum_p w_p^e \zeta_p \quad (8)$$

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

3.2.3 An Experiment with System Identification Approach

We have implemented the AAM by means of the aam-api [9] coupled with OpenCV library [10]. The frame rate of the CCD camera is approximately 20 fps.

An experiment to check basic abilities of the proposed approach is conducted. For this aim, only a subject who acts the six emotions according to the FACS is assumed to be the target person. Thus, the training data for AAM and the FED is prepared only for the subject. After construction of the AAM and the FED, another sequence of the actions is observed. In the observation phase, emotion estimation process works in the real time, and the results are sent to the conceptual cognition engine in response to the facial expressions of the subject. To confirm the abilities of the approach, the subject acts the six emotions in the following order: happy, surprise, anger, disgust, sadness, and fear. The frames corresponding to the facial expressions are listed in Table 1.

TABLE 1. Frames and Acted Facial Expression.

Acted Emotional Facial Expression	Frames
Happy	51 to 105
Surprise	131 to 170
Anger	222 to 268
Disgust	310 to 346
Sadness	394 to 430
Fear	535 to 569

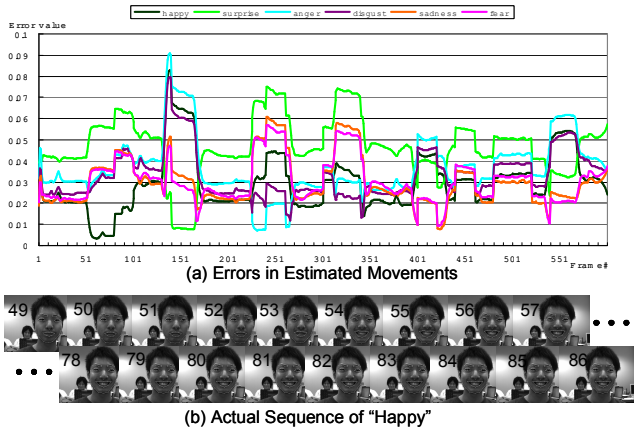


Fig.3 Result of Experiment with System Identification

The result of the experiment is depicted in Fig.3. Fig.3(a) shows the transitions of sums of the errors in Equation(5) at each frame. Therefore, the lower error value means that the corresponding emotional facial expression is observed. Comparing Fig.3 with Table 1 reveals interesting facts. The error values rapidly decrease at the beginning of each act, and in most cases, continues decreasing until the facial expression becomes the typical one that can be recognized with FACS. Fig.3(b) is the actual sequence of the facial expression expressing "Happy" from beginning to the end. This implies that the proposed method is able to detect the very beginning of the facial expressions. Although some emotions, disgust and happy, and fear and sadness, seem hard to distinguished by the proposed method, the other emotions are detectable with the method. The undistinguished emotional expressions are sometimes very hard to identify even by the human when forced to identify only from the facial expression. We are planning to merge and/or analyze the information from the other modalities, i.e., situation, voice, and so on, to overcome this issue.

3.3 Facial Expression Map Approach

Another approach to facial expression analysis is based on the positions of the feature points. For our aim, understanding the emotional states of the general public is indispensable. The previous approach provides precise and somewhat sensitive analysis of the facial expressions, but so far, it exhibits a defect in estimating the emotions of the general public. Namely, the performance is not so good when applying it to unknown persons. The preciseness of the estimation and the applicability to the general public might be stood with a tradeoff relationship.

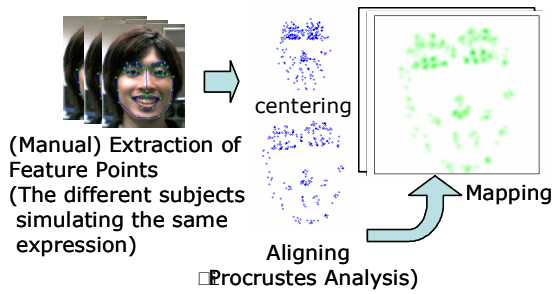


Fig. 4 Construction of Facial Expression Maps

The approach proposed in this subsection tackles with the applicability to the general public. From this view point, the approach is much alike with those methods that rely on the FACS, but differs in that it tries to estimate not only the exhibited emotions but the strengths of them. Moreover, the proposed method should not rely on the AUs (Action Units) in the FACS, because the target users of the system are not assumed to provide "neutral" face at the beginning of the interactions. For example, a person who start with smiling may exist.

Thus, the proposed method introduces a set of maps, i.e., Facial Expression Maps (FEMs), of the feature points. Each map represents the strengths of the particular emotion with respect to the discretized positions of each feature point as described in the following sub-subsections. The rest of the subsection describes a construction and usage of the facial expression map, and on preliminary experiments with the facial expression maps with the results.

3.3.1 Construction of Facial Expression Map

The facial expression map corresponds to the FED within this approach. A construction of the facial expression map requires following three assumptions: (1) positions where the facial feature points for certain emotional expressions are highly distributed have higher strengths of that emotion, (2) the biggest facial expressions are observed when a person acts their emotion on his/her face, and (3) the bigger the emotion strength is, the bigger facial expression would be observed. The basic idea underlying here is that the distribution of the feature points obtained from simulated facial expressions, contains characteristic positions of the feature points for certain emotions.

As shown in Fig.4, the construction of the map begins with gathering a set of facial expressions of the basic six emotions and neutral (no emotion as far as possible) simulated by a group of subjects. The feature points are extracted (manually or by means of AAMs) for each expression. To eliminate translational differences between the centers of each set of the feature points, a centering technique using center of gravity is adopted. Because the expressions are not

provided with the same scale, nor rotation, we have to align them to be nearly in the same scale and rotation. The Procrustes Analysis [6] is adopted for this aim.

With the aligned feature points, the FEM is constructed. Construction of the map requires a global coordinate system and as many local coordinate systems as the number of the actual feature points gathered for it. The global coordinate system is equivalent with that of the images from the CCD camera, and acts as the coordinate systems of the facial expression maps. The local coordinate systems are temporal ones that is used only to calculate the distribution of the strengths generated from a feature point at the simulated emotional expression. The origin of the coordinate is the feature point, and its x-coordinate is toward the direction of the movement observed in the simulated expression.

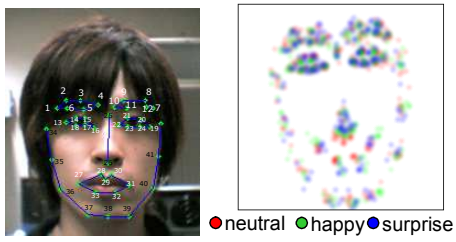
The distribution of the strength by a single feature point is provided by the following equations:

$$E^e{}_k(x_i, y_j) = \begin{cases} \frac{1}{N}\exp(-x_i{}^4)\exp(-y_j{}^2/2), (x_i < 0) \\ \frac{1}{N}\exp(-x_i)\exp(-y_j{}^2/2), (x_i \geq 0) \end{cases} \quad (9)$$

where, N is the number of the expressions used to construct the map, e implies corresponding emotion, $k \in \{1, \cdots, N\}$, and (x_i, y_j) are the points in the local coordinate system for a feature point. These distributions are added on the global coordinate system after making coordinate conversions. In the global coordinate system, $L \times M$ grids are prepared as the discrete coordinate system.

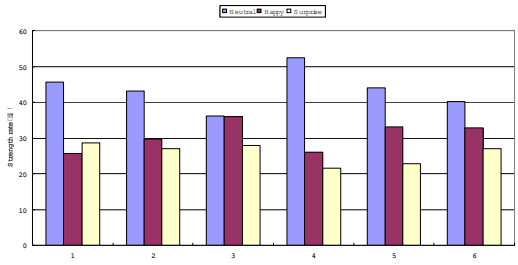
3.3.2 Facial Expression Analysis with FEM

Each grid of the FEMs contains strength of the position with respect to the corresponding emotion. All the comparer has to do is to align the observed feature points with the average feature points among those used in constructing the FEM. This is also done by means of the Procrustes Analysis. Then, the strength of the corresponding position in the map is a relevance of the observed feature point with the emotion. Note that all observed points are tried on the FEMs of all kinds, i.e., maps of the basic six emotions and neutral. The obtained strengths are to be processed by taking the positional factors and the relations among the emotions into account, to estimate the expressed emotions. This, however, remains as a future work. For now, we simply add all the strengths for the estimation, and just considering by rate to estimate what emotions are expressed.

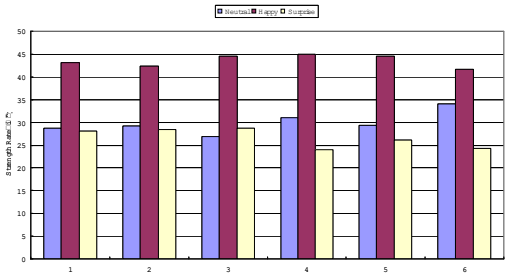


(a) assigned FPs (b) obtained maps

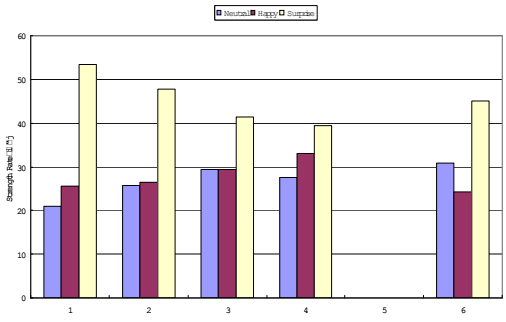
3.3.3 Preliminary Experiments on FEMs



(a) Compared with "Neutral"



(b) Compared with "Happy"



(c) Compared with "Surprise"

Fig.5 Feature Points and Superposed Maps

Fig. 6 Results of Preliminary Experiment for Test Data

Two preliminary experiments are conducted to show the possibilities of the FEM approach. For this experiments, the maps of happy, surprise, and neutral are constructed with the simulated facial expressions of (6, 5, 6) subjects respectively. Then, two trials are examined: analysis of the expressions used in the construction of the maps, and analysis of the expressions by unknown subjects. All expressions, here, are static images of the simulated expressions.

The configuration of the feature points is depicted in Fig.5(a). 31 points (with white letters) out of 42 are used as the feature points, and the rest of the points are used for the AAM together with the 31 points.

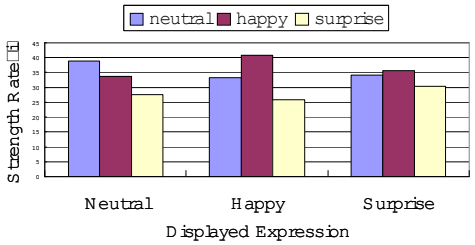
Fig.5(b) superposes the obtained points for the emotions to show the differences. In this preliminary experiments, the distributions of the strengths are given by the Gaussian, alternative to Equation (9), around the feature points to ease the computation. As shown the figure, maps around mouth significantly differ among the emotions. To know what emotion is displayed, an equation to calculate the rates is used as a strength rate:

$$E_e = \frac{\sum_i map_e(x_i, y_i)}{\sum_e \sum_i map_e(x_i, y_i)} \times 100 \qquad (10)$$

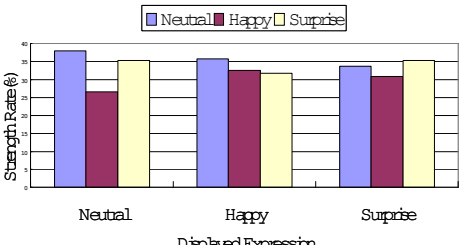
where, $map_e(x_i, y_i)$ is a strength with respect to an observed feature point.

Fig.6 shows the results of the first experiments, i.e., testing maps with the expressions used for their constructions. As shown in the graphs, the every displayed emotional expressions show high strength rates. This means that at least the displayed expressions are used in the construction phase, we can detect the emotional expression correctly.

Fig.7 shows the results obtained with unknown subjects. For Subject A who is a male student, emotional expressions other than "surprise" are detectable, while for Subject B who is a female student, emotional expressions other than "happy" are detectable. The contrast between the rates against Subject B is lower than that of A. This implies that the results are influenced by the number of the expressions used in the construction phase. In this experiment, the expressions of only one female student is used for the constructions of the maps. From this evidence, we are considering that some categorizations to the users/subjects are required in the constructions of the maps and in their usage. We are planning to introduce physiognomy for this issue.



(a) Student A (male)



(b) Student B (female)

Fig.7 Results for Unknown Subjects

4 Conclusion

This paper introduces current states of the facial expression estimation methods toward the MK Project. We have introduced three characteristics to be achieved by the autonomous facial expression analysis for intelligent HCI systems, and two approaches to realize the three characteristics are introduced. The results of the (preliminary) experiments show that both approach have potential to estimate emotions of the user from their facial expressions. We are planning to merge these two approaches to make them work complementary to estimate emotions of the user.

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*, 10(5), 2006, pp. 1536-1543.

[2] J. Hakura, et. al., Facial Expression Recognition and Synthesis for Virtual Miyazawa Kenji System, *WSEAS Transactions on Circuits and Systems*, 3(6), 2007, 288-295.

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

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

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

[6] 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.

[7] J. Velásquez, Modeling Emotions and Other Motivations in Synthetic Agents. *Proceedings of AAAI-97,* 1997.

[8] 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.

[9] Stegmann, M. B., et.al., FAME – a flexible appearance modeling environment, *IEEE Transactions on Medical Imaging*, 22(10), 2003, 1319-1331.