Facial Expression Representation Based on Timing Structures in Faces

Masahiro Nishiyama, Hiroaki Kawashima, Takatsugu Hirayama,
and Takashi Matsuyama

Kyoto University, Yoshida-Honmachi Sakyo, Kyoto 6068501, Japan
{nisiyama, hiroaki, hirayama, tm}@vision.kuee.kyoto-u.ac.jp

Abstract. This paper presents a method for interpreting facial expressions based on temporal structures among partial movements in facial image sequences. To extract the structures, we propose a novel facial expression representation, which we call a *facial score*, similar to a musical score. The facial score enables us to describe facial expressions as spatio-temporal combinations of temporal intervals; each interval represents a simple motion pattern with the beginning and ending times of the motion. Thus, we can classify fine-grained expressions from multivariate distributions of temporal differences between the intervals in the score. In this paper, we provide a method to obtain the score automatically from input images using bottom-up clustering of dynamics. We evaluate the efficiency of facial scores by comparing the temporal structure of intentional smiles with that of spontaneous smiles.

1 Introduction

Facial expression plays an important role in our communication; for instance, it can nonverbally express emotions and intentions to others. Much progress has been made to build computer systems that recognize facial expression for human interfaces. However, these systems have problems; they don't use enough dynamic information in recognition, and the classification of facial expression relies on a fundamental category based on emotions. Most previous systems describe facial expression based on action units (AUs) of the Facial Action Coding System (FACS) developed by Ekman and Friesen [13]. An AU is defined as the smallest unit of facial movement that is anatomically independent and visually distinctive. FACS is a method for describing facial expression on the basis of the combination of AUs. FACS, however, has a major weakness; there is no time component of the description [6]. Furthermore, there may be facial motion that AUs cannot express because they are heuristic motion patterns classified by human. It is also important to decide what categories of facial expression are appropriate as the outputs of facial recognition. Most previous systems categorize facial expression into one of six basic categories (happiness, surprise, fear, anger, disgust, and sadness) [6]. In human communication, however, facial expression is classified into one of the more fine-grained categories by subtle dynamic changes that are observed in facial components: the variety of changes and the timing of

W. Zhao, S. Gong, and X. Tang (Eds.): AMFG 2005, LNCS 3723, pp. 140–154, 2005.

changes. To capture the subtlety of human emotion and intention, automated recognition of subtle dynamic changes in facial expression is needed.

In this paper we assume that (1) dynamic movement of each facial component (facial part) yields changes of facial expression, and that (2) movement of facial parts is expressed based on temporal intervals. We define the intervals as temporal ranges of monotonically changing events that have beginning times, ending times, and labels of motion patterns (modes) as attributes. We provide a framework for recognizing facial expression in detail based on *timing structures*, which are defined as temporal relations among the beginning and ending times of multiple intervals. To extract the timing structures, we propose a novel facial expression representation, which we call a *facial score*. The score is similar to a musical score, which describes the timing of notes in music. Using the score, we can describe facial expressions as spatio-temporal combination of the intervals.

It is important to decide what the definition of modes is in the interval-based description. Whereas AUs are suitable to distinguish emotional facial expression, they sometimes do not preserve sufficient dynamic information (e.g., time-varying patterns) of facial actions. In this paper, we take another approach that determines a set of modes from statistical analysis and describes facial actions based on generative models. This approach extracts modes that have enough dynamic information from the viewpoint of pattern generation, and provides a unified framework that can be used not only for facial expression analysis but for facial expression generation. We propose a bottom-up learning method to find modes from captured real data. In this method, each mode is modeled by a dynamical system that has an ability of generating simple patterns, and the modes are extracted from clustering analysis based on the distances between dynamical systems (see Section 3.3 and 4.2 for details).

In summary, the facial score is characterized as follows:

- It enables us to describe timing structures in faces based on temporal intervals.
- It enables us to use motion patterns extracted from training data in a bottom-up manner as modes of intervals.

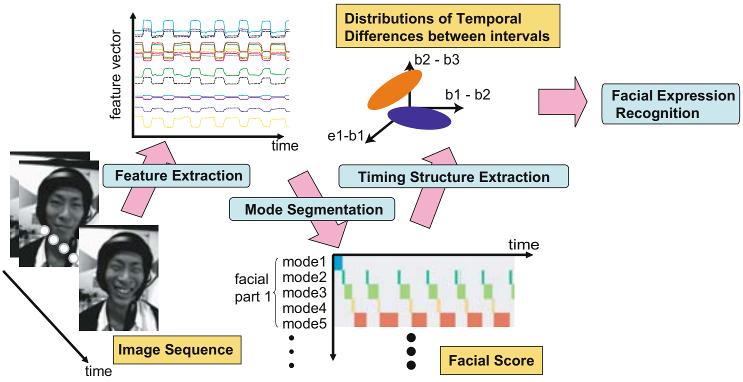


Fig. 1. The overall flow of facial expression recognition using the facial score

Figure 1 depicts the overall flow of facial expression recognition using the facial score: (1) we extract a series of feature vectors that characterize facial expression from a sequence of facial images, (2) we partition the series of feature vectors and extract the modes simultaneously to obtain a facial score, and (3) we extract timing structures from the facial score, which contribute to recognition of the facial expression. Automation of the above process provides for applications in recognizing facial expression, and therefore allows computers to learn to recognize facial expression in detail.

The goal of this paper is to propose a method for automatically obtaining the facial score and to evaluate the efficiency of the facial score for facial expression recognition. We compare the timing structure of intentional smiles with that of spontaneous smiles for the evaluation; in human communication it makes sense to make a distinction between the two smiles, but most previous computer systems have classified these smiles into the same category.

In Section 2, related works are described. In Section 3, facial scores are introduced as representations that describe timing structures in faces. In Section 4, we describe a method for automatically obtaining the facial score from input sequences of facial images. In Section 5, we obtain facial scores automatically from captured real data including intentional and spontaneous smiles, and evaluate the efficiency of the facial scores by the separability between the two smiles. Finally, in Section 6 we conclude our work.

2 Related Works

In psychological experiments, evaluation by playing back facial expressions on videotape to subjects has suggested the following knowledge of dynamic aspects of facial movement. Bassili video-recorded the face that was covered with black makeup and numerous white spots, and found that it is possible to distinguish facial expression to a certain degree of accuracy merely from motion of the white spots by playing back the video [2]. As a study concentrating on a more specific part of facial motion, Koyama, et al. created CG animations with the temporal relation between eye and mouth movement controlled, and showed laughter can be classified into pleasant, unpleasant, and sociable types based on the temporal difference [9]. As a study of analyzing solitary and social smiles, Schmidt, et al. indicated temporally consistent lip movement patterns based on the evaluation of the relationship between maximum velocity and amplitude [11]. Hence, the importance of dynamic aspect in facial expression has been emphasized by many studies. However, an appropriate representation that maintains spatio-temporal structures in facial actions is still under study.

3 Facial Scores

3.1 Facial Scores Definition

A facial score is a representation that describes motion patterns of each facial component and temporal relations between the movement. In this paper we define the following notations:

Facial parts and Facial Part Sets: Facial parts represent isolable facial components. We define facial part sets as $\mathcal{P} = \left\{P_1, ..., P_{N_p}\right\}$ where N_p is the number of facial parts described by facial scores. For instance, elements of facial part sets include mouths, right eyes, left eyes, right eyebrows, and left eyebrows.

Modes and Mode Sets: Modes represent monotonically changing events. We define mode sets as $\mathcal{M}^{(a)} = \left\{M_1^{(a)}, ..., M_{N_{m_a}}^{(a)}\right\}$ where N_{m_a} is the number of modes of a facial part P_a $(a \in \{1, ..., N_p\})$. For instance, elements of mode sets of a mouth part include "opening", "remain open", "closing", and "remain closed".

Intervals and Interval Sets: Intervals represent temporal ranges of modes. We define interval sets as $\mathcal{I}^{(a)} = \left\{I_1^{(a)}, ..., I_{N_{k_a}}^{(a)}\right\}$ where N_{k_a} is the number of intervals into which time series data of a facial part P_a is segmented. Intervals $I_k^{(a)}$ $(k \in \{1, ..., N_{k_a}\})$ have beginning times $b_k^{(a)} \in \{1, ..., T\}$, ending times $e_k^{(a)} \in \{1, ..., T\}$, and labels of modes representing the events $m_k^{(a)} \in \mathcal{M}^{(a)}$ as attributes where T is the number of time series data of a facial part P_a.

Facial Scores: We define a facial score as a set of interval sets of all facial parts $\{\mathcal{I}^{(1)}, ..., \mathcal{I}^{(N_p)}\}$. Figure 2 shows a conceptual figure of a facial score. The vertical axis represents modes of facial parts, and the horizontal axis represents time. The transition of the motion of each facial part is described based on intervals along the temporal axis. In each facial part of the figure, intervals of various colors represent various modes. Thus, the facial score can describe timing structures among motions of facial parts.

3.2 Facial Parts in Facial Scores

To recognize facial expression based on timing structures, we treat the two areas where their movements occur independently as different facial parts. Ekman, *et al.* have revealed that the difference in the facial appearance of basic emotions (happiness, surprise, fear, anger, disgust, and sadness) results from the combination of the three facial areas (around the eyebrows, eyes, and mouth) where their movements can be observed individually in appearance [5]. We use these

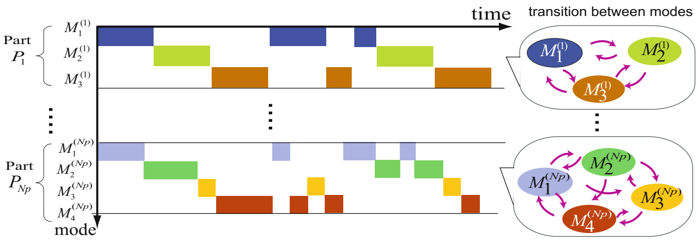


Fig. 2. Facial scores. The vertical axis represents modes of facial parts, and the horizontal axis represents time. The transition of the motion of each facial part is described based on intervals along the temporal axis.

three areas, and furthermore treat areas around the eyebrows and eyes on the left and right as different facial parts because the asymmetric movements of each eyebrow and eye can be observed in real facial expression.

It is important to select useful features that can express subtle changes of movements in the five facial areas. This paper defines feature vectors as coordinates of feature points shown in Figure 5 (a), which can extract information of movement directly. We consider that transient features such as furrows also provide effective information in recognition of subtle facial expression, and that changes of the feature points can represent them indirectly; for instance, movement of feature points on the nose implies nasolabial furrows.

Therefore, we define elements of facial part sets $\mathcal{P}$ as right eyebrow, left eyebrow, right eye, left eye, nose, and mouth. A feature vector $z^{(a)}$ of a facial part P_a is represented by the following $2n_{p_a}$-dimensional column vector:

$$z^{(a)} = \left(x_1^{(a)},\ y_1^{(a)}, ..., x_{n_{p_a}}^{(a)},\ y_{n_{p_a}}^{(a)}\right)^\top , \tag{1}$$

where n_{p_a} is the number of feature points of a facial part P_a, and let $\left(x_p^{(a)},\ y_p^{(a)}\right)$ be coordinates of a feature point number $p \in \{1, ..., n_{p_a}\}$.

3.3 Modes in Facial Scores

As we defined in Section 3.1, each complex movement of a facial part is composed of simple motion categories, which we call modes. Therefore, a movement can be partitioned into a sequence of temporal intervals by modes.

Modes are classified into two large categories by the velocity of feature vectors: stationary poses and dynamic movements. For the modes with movement, we use monotonic motions as the lowest-level representation, whereas humans sometimes classify a cyclic motion as one category. Therefore, our facial score represents a cyclic motion as a sequence of monotonic motions. For example, the open and close action of the mouth is represented as the following sequence of four modes: “opening”, “remain open”, “closing”, and “remain closed”.

AUs used in FACS are the most common units to describe facial movements. Although AUs are suitable to distinguish emotional facial expressions by their combinations, we do not use AUs as the modes in our facial scores for two reasons. First, a method of AU tracking is still a challenging research topic for computer vision. Second, AUs sometimes do not maintain sufficient dynamic information in facial actions. As a result, AU-based CG animation systems sometimes generate unnatural facial actions. In contrast, our approach takes a bottom-up learning method to find modes rather than using predefined motion categories, as we described in Section 1. That is, all the modes are extracted by the clustering of dynamics from captured real data, as we will see in Section 4.2. For a generative model of simple dynamics in each mode, we use a first-order linear dynamical system. The dynamics of the mode $M_i^{(a)}$ $(i \in \{1, ..., N_{m_a}\})$ in a facial part P_a is represented by the following notation:

$$z_t^{(a)} = F^{(a,\ i)} z_{t-1}^{(a)} + f^{(a,\ i)} + \omega_t^{(a,\ i)}, \tag{2}$$

where $z_t^{(a)}$ is a feature vector at time t, $F^{(a,\ i)}$ is a transition matrix, which differs from other modes' matrices, $f^{(a,\ i)}$ is a bias term, $\omega^{(a,\ i)}$ is a process noise of the system that has a multivariate Gaussian distribution with mean vector 0 and covariance matrix $Q^{(a,\ i)}$.

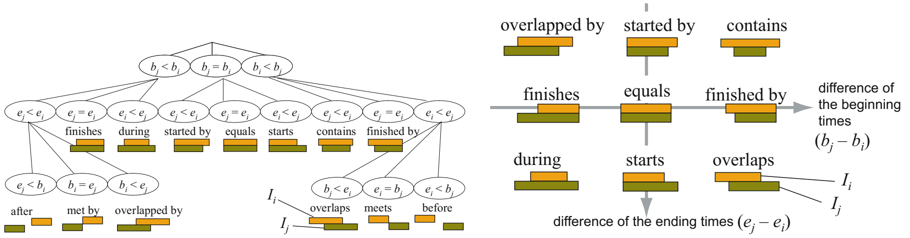
As a result, each motion transition in each facial part is described based on the transition of linear dynamical systems, which is similar to a switching linear dynamical system [3,8]. Therefore, the proposed model can be considered as a concurrent process of multiple switching linear dynamical systems. We currently do not model the transition probability between modes to reduce the model parameters; however, the transition probability will work as constraints during a mode segmentation process, and can be introduced if specific mode transition patterns appear frequently.

Given a sequence of feature vectors, we find a rough segmentation using zero-crossing points of the velocity as the initialization of the method. Then, we merge the nearest dynamical system pairs iteratively based on agglomerative hierarchical clustering. A linear dynamical system, in general, can generate not only monotonic motions but cyclic or oscillating motions. To extract only the monotonic motions, we propose a method to provide a constraint on eigenvalues of the transition matrices. We will describe the details of the identification and clustering algorithms in Section 4.2.

3.4 Timing Structures in Facial Scores

Using facial scores defined in the previous sections, we can represent temporal relations among motions in facial parts; we refer to the relation as timing structures of the face. In this section, we describe a method to represent and extract timing structures from a facial score.

Representation of Timing Structures: Figure 3(a) shows 13 categories of temporal relations between two intervals I_i and I_j [1,10]. We can classify the relations of the two intervals based on the temporal order of four times b_i, b_j, e_i and e_j, where $b_i(b_j)$ and $e_i(e_j)$ represent the beginning and ending times of the interval $I_i(I_j)$, respectively. Although these categories enable us to represent temporal structures among multiple events, such as overlaps between two intervals, they are insufficient for us to describe the difference of timing structures in facial expressions. We need to concentrate on not only temporal order of events but scales and degree of temporal differences among beginning and ending times of multiple intervals. In this paper, we extend the 13 categories based on multivariate distributions of real-valued variables. Using temporal differences between beginning and ending times, we can represent the first-order timing structure of two intervals as four distributions $H(b_j - b_i), H(e_j - e_i), H(b_j - e_i)$ and $H(e_j - b_i)$, where $H(r)$ is a one-dimensional distribution of variable $r \in \mathrm{R}$. We can also represent the second-order timing structure as six distributions $H(b_j - b_i, e_j - e_i), H(b_j - b_i, b_j - e_i), H(b_j - b_i, e_j - b_i), H(e_j - e_i, b_j - e_i), H(e_j - e_i, e_j - b_i)$ and $H(b_j - e_i, e_j - b_i)$, where $H(r_1,\ r_2)$ is a two-dimensional distribution of variables $r_1,\ r_2 \in \mathrm{R}$. Figure 3(b) shows the example of distribution $H(b_j - b_i,\ e_j - e_i)$,



(a) Temporal relations between two intervals (b) Two-dimensional temporal distribution

Fig. 3. (a) An example of two-dimensional distributions of temporal differences between two intervals. The temporal order of beginning and ending times provides 13 relations of the two intervals. (b) The horizontal and vertical axes denote the difference between beginning times $b_j - b_i$ and the difference between ending times $e_j - e_i$ of the two intervals (I_i and I_j), respectively.

where the horizontal and vertical axes represent the difference between the beginning times and the difference between the ending times, respectively. Representations of high-order timing structures become a set of high-dimensional distributions in the same manner. To represent timing structures among more than three intervals, for example the first-order timing structure of three intervals I_i, I_j and I_k, we need 12 one-dimensional distributions $H(b_j - b_i), H(b_k - b_j)$, and so on.

Extraction of Timing Structures from Facial Scores: The selection of interval combinations is necessary for calculating the distributions that are described in the previous paragraphs when we make use of the timing structures for facial expression analysis and recognition. In our experiments in Section 5, we selected the combinations based on the following methods. First, we find combinations of the intervals that belong to each facial part based on temporal distances. The interval in a facial part P_b $(b \neq a, b \in \{1, ..., N_p\})$ that has the nearest distance from an interval $I_k^{(a)}$ in a facial part P_a is calculated as $I_{l^*}^{(b)} (l^* = \arg\min_l \mathsf{IntervalDist}(I_k^{(a)},\ I_l^{(b)}))$, where $\mathsf{IntervalDist}$ is a distance between two intervals that is defined as follows:

$$\mathsf{IntervalDist}(I_k^{(a)},\ I_l^{(b)}) = |b_k^{(a)} - b_l^{(b)}| + |e_k^{(a)} - e_l^{(b)}|. \tag{3}$$

Second, we represent the timing structure as two-dimensional distributions. If there are clusters in the calculated distributions, we can define successfully more subtle categories of facial expressions than basic emotional facial expressions.

4 Automatic Acquisition of Facial Scores

In this section, we describe a method for automatically obtaining facial scores with facial image sequences as the inputs.

4.1 Facial Feature Extraction

We track feature points in facial image sequences using Active Appearance Models (AAM) [4]. An AAM contains a statistical model of correlations between shape and grey-level appearance variation. The model can be matched to a target image rapidly and robustly.

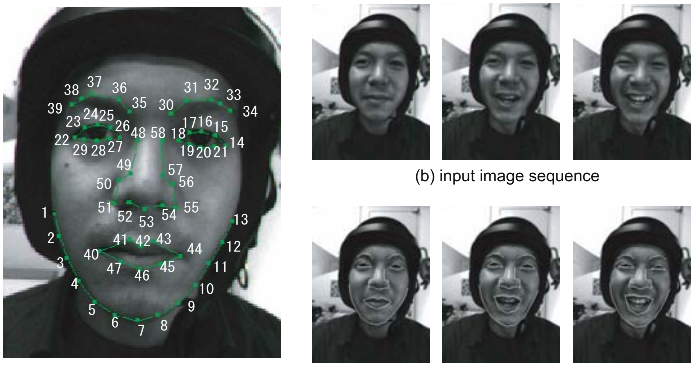
To build the model, we require a training set of images marked with feature points. Figure 4 (a) shows an example of a face image labeled with 58 feature points. Let s be a shape vector that represents the coordinate value of feature points. Let g be a grey-level vector that represents the intensity information from the shape-normalized image over the region covered with the mean shape. In the first step, the method applies principal component analysis (PCA) to the data. Any example image can then be approximated using:

$$s = \bar{s} + U_s c_s \ , \ g = \bar{g} + U_g c_g, \tag{4}$$

where $\bar{s}$ and $\bar{g}$ are the corresponding sample mean vectors, U_s and U_g are matrices of column eigenvectors of the shape and grey-level, and c_s and c_g are vectors of shape and grey-level parameters, respectively. In the second step, because there may be correlations between the shape and grey-level variation, the method concatenates the vectors c_s and c_g, applies PCA, and obtains a model of the form

$$\begin{bmatrix} W_s c_s \\ c_g \end{bmatrix} = c = \begin{bmatrix} V_s \\ V_g \end{bmatrix} d = V d, \tag{5}$$

where W_s is a diagonal matrix of weights for each shape parameter, allowing for the difference in units between the shape and grey-level models, V is a matrix of column eigenvectors, and d is a vector of appearance parameters controlling both the shape and grey-levels of the model.



(b) input image sequence

(a) feature points to track

(c) tracked feature points using active appearance model

Fig. 4. (a) A training image to build active appearance models. (b) Part of a captured face image sequence. (c) Part of a face image sequence with tracked feature points.

Note that the linear nature of the model allows us to express the shape vector s and grey-level vector g directly as functions of d:

$$s = \bar{s} + U_s W_s^{-1} V_s d \ , \ g = \bar{g} + U_g V_g d. \tag{6}$$

An example image can be synthesized for a given d by generating the shape-free grey-level image from the vector g and warping it using the feature points described by s. During a training phase we learn the relationship between model parameter displacements and the residual errors induced between a training image and a synthesized image.

The matching process for tracking the feature points is provided as an optimization problem in which we minimize the difference between a target image and an image synthesized by the model.

4.2 Modes Extraction

As we postulated in Section 3.3, each mode in the facial expression score is represented by a different linear dynamical system. In this section, we describe a method to find a set of modes that corresponds to a set of dynamical systems. This algorithm is applied to each facial part independently.

Although there are several approaches to find dynamics in training sequences, we propose a bottom-up clustering method to extract modes based on an agglomerative hierarchical clustering approach described in [7]. The method provides useful interfaces such as dendrograms to determine the number of clusters.

First, we introduce a constrained system identification method that restricts an upper bound of eigenvalues in the transition matrix in Equation (2). The method enables us to find a set of modes that represent only monotonic dynamics. Then, we introduce an agglomerative hierarchical clustering of dynamical systems with the definition of distance between two dynamical systems. This algorithm also merges two interval sets that are labeled by the same dynamical system in each iteration. Thus, the clustering methods solve two problems simultaneously: temporal segmentation and parameter estimation.

Constrained System Identification: The parameter estimation of a transition matrix $F^{(a,i)}$ from a sequence of feature vectors $z_1^{(a,i)}, .., z_T^{(a,i)}$ in a facial part P_a becomes an error minimization problem; that is, minimizing squared prediction error vectors during the temporal interval $[1, T]$ that is represented by the mode $M_i^{(a)}$. For convenience, we drop index a, which identifies a facial part, in the remaining of this section because the following clustering method is applied to each part independently.

The key idea to estimate monotonic dynamics is the method to constrain on eigenvalues. If all the eigenvalues are lower than 1, the dynamical system changes state in a monotonic manner (i.e., cyclic or oscillation will not occur and the state converges to a certain value). Using the notation $Z_0^{(i)} = [z_1^{(i)}, ..., z_{T-1}^{(i)}]$ and $Z_1^{(i)} = [z_2^{(i)}, ..., z_T^{(i)}]$, we can estimate matrix $F^{(i)}$ by the following equation:

$$F^{(i)*} = \arg\min_{F^{(i)}} ||F^{(i)} Z_0^{(i)} - Z_1^{(i)}||^2 = \lim_{\delta^2 \to 0} Z_1^{(i)} Z_0^{(i)\top} (Z_0^{(i)} Z_0^{(i)\top} + \delta^2 I)^{-1}, \tag{7}$$

where I is a $2n_{p_a} \times 2n_{p_a}$ unit matrix and δ is a positive real value. Using Gershgorin's theorem in linear algebra, we can determine the upper bound of eigenvalues in a matrix from the elements of the matrix. Therefore, we use a nonzero value for δ that controls the scale of values in the matrix; that is, we stop the limit in the Equation (7) before $Z_0^{(i)\top}(Z_0^{(i)}Z_0^{(i)\top} + \delta^2 I)^{-1}$ converges to a pseudo-inverse matrix of $Z_0^{(i)}$.

Clustering of Dynamics: The clustering algorithm of dynamics (modes) is initialized by a segmentation that partitions the training sequence into motion and stationary pose intervals, which we call the initial interval set. To calculate the initial interval set, we simply divide the training sequence by zero-crossing points of feature velocity (i.e., the first-order difference of feature vectors). In the first step of the algorithm, one dynamical system is identified from each interval in the initial interval set. Then, we calculate the distances for all the dynamical system pairs based on the distance definition in the next paragraph. In the second step, the nearest dynamical systems are merged iteratively based on an agglomerative hierarchical clustering (see Algorithm 1 in Appendix for details). Finally, all the modes are merged to one mode. We determine the number of the modes manually using the obtained dendrogram (i.e., the tree structure that provides the history of the total distance change).

Distance Between Dynamical Systems: We define the distance between two dynamical systems (modes) based on a cross check of the prediction errors between the two modes. In the following equation, we use the notation $z_{t-1}^{(i)}$ and $z_t^{(i)}$, which means that the adjacent feature vectors z_{t-1} and z_t belong to an interval that is represented by mode M_i. The prediction from the vector $z_{t-1}^{(i)}$ to the feature vector of time t by the dynamics of the mode M_j becomes $F^{(j)}z_{t-1}^{(i)}$. Thus, we can calculate the prediction error from $z_t^{(i)}$ as $E_t^{(i|j)} = F^{(j)}z_{t-1}^{(i)} + f^{(j)} - z_t^{(i)}$. Calculating this prediction error for all the adjacent feature vectors in the interval set $\mathcal{I}_i$, which is represented by the mode M_i, we can define the prediction error from M_j to M_i as the following equation:

$$E\left(M_i || M_j\right) = \frac{1}{C} \sum_{I_k \in \mathcal{I}_i} \sum_{t=b_k}^{e_k} ({E_t^{(i|j)}}^2 - {E_t^{(i|i)}}^2), \tag{8}$$

where C is the total interval length of the intervals in the set $\mathcal{I}_i$, which normalizes the sum of prediction error in a time axis. For the distance definition between two modes, we take the average

$$\mathsf{Dist}\left(M_i,\ M_j\right) = \left\{E\left(M_i||M_j\right) + E\left(M_j||M_i\right)\right\}/2, \tag{9}$$

because the two prediction errors, from M_i to M_j and from M_j to M_i, are asymmetric.

5 Experimental Evaluations

We evaluated the efficiency of our representation for a separation of intentional smiles from spontaneous smiles using obtained facial scores from captured data.

Video Capturing: Intentional and spontaneous smiles of four subjects were captured in 240 × 320 at 60 fps as the input image sequences. We used a camera system that was composed of a helmet and a camera fixed in front of the helmet to focus on the analysis of front faces. The camera system enabled us to capture front face images even if head motion occurred. The subjects were instructed to begin with a neutral expression, make a smile, and return to a neutral expression again. Intentional smiles were captured by instructing the subjects to force a smile. Spontaneous smiles were captured by making the subjects laugh. The subjects were instructed to make either smile iteratively in capturing one sequence, so that no sequences included both smiles. Figure 4 (b) shows part of a captured face image sequence.

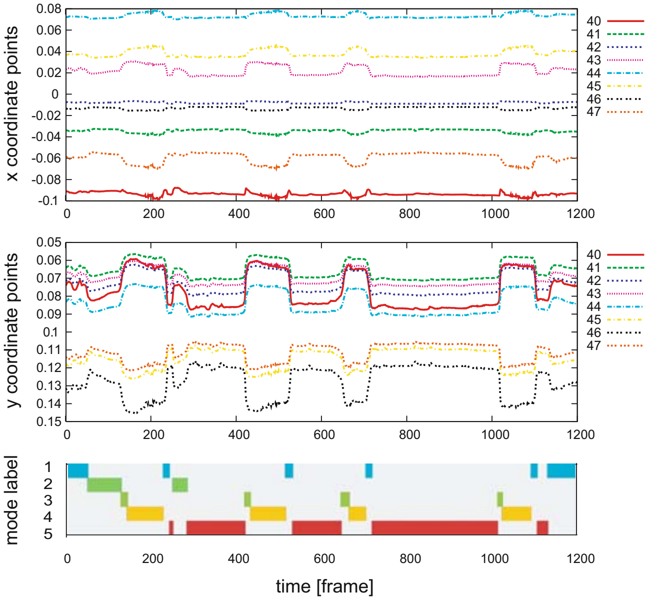


Fig. 5. The correspondence of the mouth part of an obtained facial score from spontaneous smiles with the feature vector series. The vertical axes of the top, the middle and the bottom subfigures represent x-coordinates of feature points, y-coordinates of feature points and modes respectively, and the horizontal axes of each subfigure represent time. The numbers of legends in the top and middle correspond to numbers that represent labels of feature points in Figure 4 (a). For example, the mode 4 and 5 represent "remain open" and "remain closed", respectively.

Automatic Acquisition of Facial Scores: Feature points in the captured face image sequences were tracked using the method in Section 4.1[1]. The number

[1] Feature points were tracked using the AAM-API that Stegmann (Technical University of Denmark) developed [12].

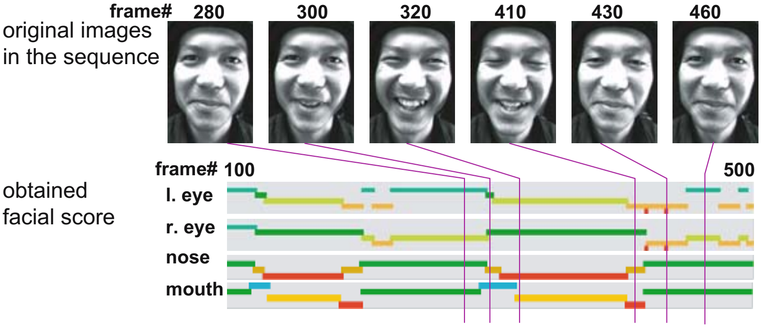


Fig. 6. An example of obtained facial scores from intentional smiles (left and right eyebrows are omitted)

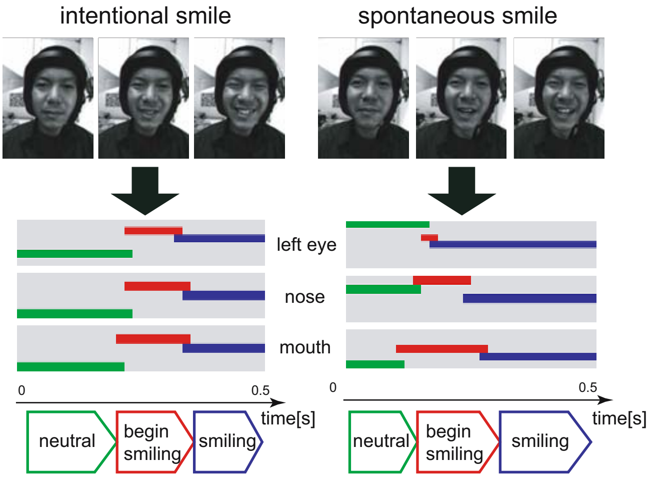
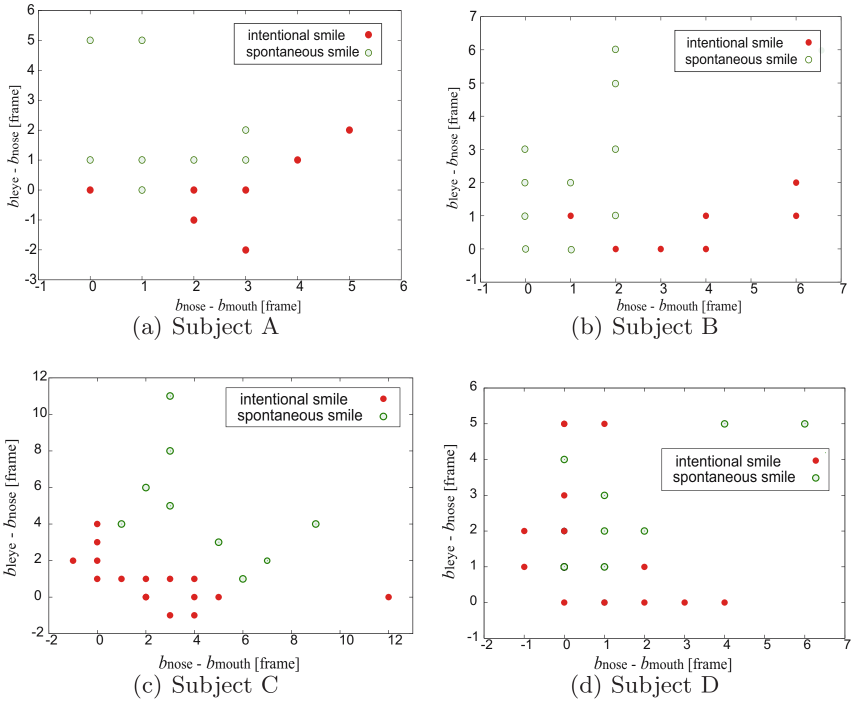


Fig. 7. Comparison of the two facial scores obtained from intentional and spontaneous smiles

of feature points used in the AAM was set to 5 on each eyebrow, 8 on each eye, 11 on the nose, 8 on the mouth, and 13 on the jawline (refer to Figure 4 (a)). Although the jawline was not represented as one of the facial parts, it was used for improving tracking accuracy. Therefore, feature vectors were obtained whose dimensions for each eyebrow, each eye, the nose, and the mouth were 10, 16, 22 and 16 respectively. Figure 4 (c) shows part of a face image sequence with tracked feature points; the frames correspond to the images shown in Figure 4 (b). Comparison of the corresponding images demonstrates extremely precise detection of feature points in changes of facial expression.

The obtained feature vectors of each facial part were segmented into modes using the method in Section 4.2. Consequently, facial scores of intentional and spontaneous smiles were acquired. Figure 5 shows the correspondence of the mouth part of an obtained facial score from spontaneous smiles with the feature



(a) Subject A (b) Subject B

(c) Subject C (d) Subject D

Fig. 8. The two-dimensional distribution that represents the timing structures of the beginning of intentional and spontaneous smiles. The horizontal axis denotes the difference between the beginning times of the nose and mouth $b_{nose} - b_{mouth}$, and the vertical axis denotes the difference between the beginning times of the left eye and nose $b_{leye} - b_{nose}$.

vector series. The vertical axes of the top, the middle and the bottom subfigures represent x-coordinates of feature points, y-coordinates of feature points and modes respectively, and the horizontal axes of each subfigure represent time. Figure 6 shows an example of obtained facial scores from intentional smiles, and the correspondence it with captured image data. These figures demonstrate that movement of smiles can be segmented into the following different modes: “neutral”, “begin smiling”, “smiling”, and “end smiling”.

Comparison of Timing Structures in Intentional and Spontaneous Smiles: As an example of comparison of timing structures in intentional and spontaneous smiles, we concentrated on a mode “begin smiling” and examined temporal relations between the beginning and ending times of the mouth, nose and left eye modes (see Figure 7). 20 samples of each smile were prepared. We used a two-dimensional distribution $H(b_{nose} - b_{mouth}, b_{leye} - b_{nose})$, which separated the two smiles with the highest efficiency, where b_{mouth}, b_{nose}, and b_{leye} are the beginning times of the mouth, nose, and left eye, respectively. Figure 8 shows the distributions of four subjects. We see that there are respective clusters in the distribution of the two smiles in case of subject A, B and C, but that there are not any clear clusters in case of subject D. We can find similarity between the distributions of subject A and B. On the other hand, we can find difference between distribution of subject A and C (or subject B and C). Hence, our experimental result suggests that the timing structures extracted from facial scores have individual variation, and the timing structures are effective in discrimination of the two smiles.

6 Conclusion

We proposed a facial score as a novel facial expression representation. The score describes timing structures in faces by assuming that dynamic movement of each facial part yields changes of facial expression. Using the score, we provided a framework for recognizing fine-grained facial expression categories. In our evaluation, the scores were acquired from captured real image sequences including intentional and spontaneous smiles automatically, and we confirmed that movement of facial parts was expressed based on temporal intervals. We suggested the individual variation of the timing structures extracted from facial scores and the efficiency of the timing structures for discrimination of the two smiles.

To emphasize the characteristics of the proposed representation, we focused on only timing structures in this paper. Other features of movement such as scale, speed and duration, which provide further information on recognizing facial expression, should be taken into account in practical systems. We also need to discuss specificity and generality of timing structures: some structures may exist as general features determined by physical muscle constraints, and the other may exist as subject specific features acquired as personal habits. Directions for future works are to tackle these problems and to evaluate the effectiveness of timing structures using a large number of captured sequences.

Acknowledgment. This work is in part supported by Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contract of 13224051 and 16700175.

References

1. J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.
2. J. N. Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology: Human Perception and Performance*, 4(3):373–379, 1978.
3. C. Bregler. Learning and recognizing human dynamics in video sequences. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 568–574, 1997.
4. T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance model. *Proc. European Conference on Computer Vision*, 2:484–498, 1998.
5. P. Ekman and W. V. Friesen. *Unmasking the Face.* Prentice Hall, 1975.
6. I. A. Essa and A. P. Pentland. Facial expression recognition using a dynamic model and motion energy. *Proc. IEEE Int'l Conference on Computer Vision*, pages 360–367, 1995.

7. H. Kawashima and T. Matsuyama. Hierarchical clustering of dynamical systems based on eigenvalue constraints. *Proc. Int'l Conference on Advances in Pattern Recognition (S. Singh et al. (Eds.): LNCS 3686)*, pages 229–238, 2005.
8. Y. Li, T. Wang, and H. Y. Shum. Motion texture: A two-level statistical model for character motion synthesis. *SIGGRAPH*, pages 465–472, 2002.
9. S. Nishio, K. Koyama, and T. Nakamura. Temporal differences in eye and mouth movements classifying facial expressions of smiles. *Proc. IEEE Int'l Conference on Automatic Face and Gesture Recognition*, pages 206–211, 1998.
10. C. Pinhanez and A. Bobick. Human action detection using pnf propagation of temporal constraints. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 898–904, 1998.
11. K. L. Schmidt, J. F. Cohn, and Y.-L. Tian. Signal characteristics of spontaneous facial expressions: Automatic movement in solitary and social smiles. *Biological Psychology*, 65:49–66, 2003.
12. M. B. Stegmann, B. K. Ersboll, and R. Larsen. FAME - a flexible appearance modelling environment. *Informatics and Mathematical Modelling, Technical University of Denmark*, 2003.
13. Y. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

A Clustering of Dynamical Systems

Algorithm 1 Agglomerative Hierarchical Clustering

```
for $i \leftarrow 1$ to $N$ do
  $M_i^{(a)} \leftarrow \mathsf{Identify}\left(I_i^{(a)}\right)$
end for
for all pair$\left(M_i^{(a)},\ M_j^{(a)}\right)$ where $M_i^{(a)},\ M_j^{(a)} \in \mathcal{M}^{(a)}$ do
  $\mathsf{Dist}\,(i,\ j) \leftarrow \mathsf{CalcDistance}\left(M_i^{(a)},\ M_j^{(a)}\right)$
end for
while $N \geq 2$ do
  $(i^*,\ j^*) \leftarrow \arg\min_{(i,\ j)} \mathsf{Dist}\,(i,\ j)$
  $\mathcal{I}_{i^*}^{(a)} \leftarrow \mathsf{MergeIntervals}\left(\mathcal{I}_{i^*}^{(a)},\ \mathcal{I}_{j^*}^{(a)}\right)$;   $M_{i^*}^{(a)} \leftarrow \mathsf{Identify}\left(\mathcal{I}_{i^*}^{(a)}\right)$
  erase $M_j^{(a)*}$ from $\mathcal{M}^{(a)}$;   $N \leftarrow N - 1$
  for all pair$\left(M_i^{(a)*},\ M_j^{(a)}\right)$ where $M_j^{(a)} \in \mathcal{M}^{(a)}$ do
    $\mathsf{Dist}(i^*,\ j) \leftarrow \mathsf{CalcDistance}\left(M_{i^*}^{(a)},\ M_j^{(a)}\right)$
  end for
end while
```

The clustering algorithm is applied to each facial part independently, and extracts modes (simple motion) in the facial part. Suffix (a) in $M^{(a)}$ and $I^{(a)}$ denotes an index of facial part. Identify is a constrained system identification described in Section 4.2, which estimates the mode parameters $\theta_i^{(a)} = \{F^{(a,\ i)},\ f^{(a,\ i)}\}$ from feature vectors in intervals. $\mathcal{I}_i^{(a)}$ is an interval set that comprises intervals labeled by $M_i^{(a)}$. CalcDistance calculates the distance between the two modes based on Equation (9). MergeIntervals merges two interval sets that belong to the nearest modes (dynamical systems).