# Chapter 4

# Analysis of Timing Structures in Multipart Motion of Facial Expression

In this chapter, we see how the interval-based hybrid dynamical system (interval system) can be applied to describe and analyze structured dynamic events. As we have shown in the preceding chapters, the interval system has the ability to describe dynamic events based on interval-based representation. We now apply the system to represent complex motion appeared in each facial part independently, and to analyze the dynamic structures of facial expression based on the temporal differences among beginning and ending time points of primitive motion.

## 4.1 Timing Structures in Facial Expression

### 4.1.1 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 that they don't use enough dynamic information in recognition, and the classification of facial expression relies on a fundamental category based on emotions (happiness, surprise, fear, anger, disgust, and sadness) [EP97].

Many systems developed so far describe facial expression based on "action units" (AUs) of the Facial Action Coding System (FACS) proposed by Ekman and

Friesen [EF75, TKC01]. 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 [EP97]. Moreover, there may be facial motion that AUs cannot express because AUs 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.

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 changes. This is because human facial expression is made of two mechanisms: (1) emotional expression produced by spontaneous muscular action, and (2) intentional display to convey some intention to others. To recognize the details of human emotion and intention, we believe that the analysis of the dynamic structure in facial expression is indispensable.

To realize such systems, we first assume the following points:

- Dynamic movement of each facial component (facial part) yields changes of facial expression.
- Movement of facial parts is expressed based on temporal intervals.

Based on the assumptions above, we define each interval as a temporal range that is expressed by a simple motion, where the intervals have beginning times, ending times, and labels of motion patterns, *modes*, as attributes.

We then 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.

Whereas AUs are suitable motion units to distinguish emotional facial expression, they sometimes do not preserve sufficient dynamic information (e.g., time-varying patterns) of facial actions. Here, we take another approach; that is, we determine a set of modes from statistical analysis and describe 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.

As for the generative models, we utilize the interval systems proposed in Chapter 2. As for the mode determination, we exploit a bottom-up learning method that we proposed in Chapter 3. In this method, each mode is modeled by a linear dynamical system that has an ability to generate simple patterns, and the modes are extracted from clustering analysis that we described in Section 3.3 (see Subsection 4.2.3 for details).

In summary, the facial score is that it enables us to

- Describe timing structures in faces based on temporal intervals;
- Use motion primitives (i.e., modes) extracted from training data in a bottom-up manner.

#### Facial Expression Generation and Recognition Using the Facial Scores

Figure 4.1 depicts the overall flow of facial expression generation and recognition using the facial score (the top right of Figure 4.1).

**Facial action generation.** Once a facial score is obtained, we can activate interval systems to generate facial expression video just like as playing music according to a musical sore (down arrow at the right column in Figure 4.1).

**Facial expression recognition.** The flow of the facial expression recognition is as follows:

1. We first extract a series of feature vectors that characterize facial expression from a sequence of facial images (right arrow at the bottom in Figure 4.1).
2. We then partition the series of feature vectors and extract the modes simultaneously to obtain a facial score using interval systems and their learning method (up arrow at the right column in Figure 4.1).
3. Finaly, we extract timing structures, which contribute to recognition of the facial expression, from the facial score (left arrow at the top in Figure 4.1).

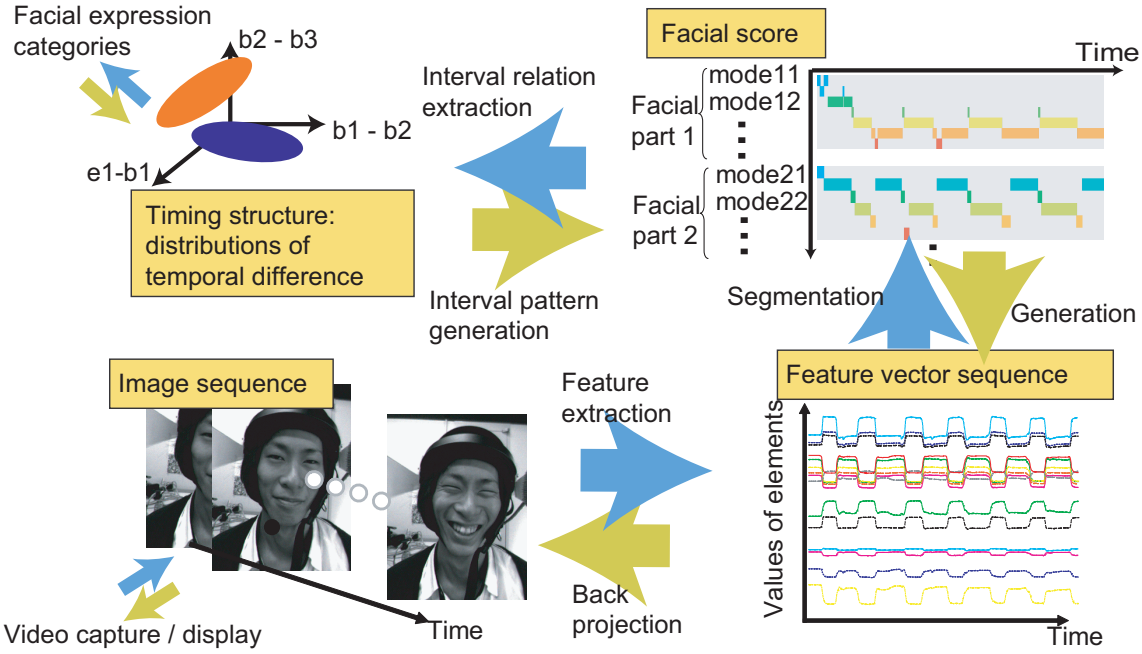



Figure 4.1: Flow of facial expression generation and recognition using the facial score.

The automation of the above process provides applications of learning, generating, and recognizing facial expression in detail using computers. The goal in this thesis is to propose a method for automatically obtaining the facial score and to evaluate the effectiveness 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 the next subsection, we describe some related work that studies dynamic properties of facial expression. In Section 4.2, we introduce facial scores as a description of timing structures in faces, and describe a method to obtain facial score. In Section 4.3, we describe a method to represent and extract timing structures from a facial score. In Section 4.4, we evaluate the effectiveness of the facial scores. In this evaluation, we first obtain facial scores automatically from captured real data including two expression categories: intentional and spontaneous smiles. Then, we examine the effectiveness of timing structure to separate these two categories of smiles. Finally, in Section 4.5 we discuss the advantage and disadvantage of the proposed representation.

### 4.1.2 Related Work

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 [Bas78].

As a study concentrating on a more specific part of facial motion, Koyama, et al. created CG animations with the temporal relations between eye and mouth movement controlled, and showed laughter can be classified into pleasant, unpleasant, and sociable types based on the temporal difference [NKN98]. 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 [SCT03]. 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.

## 4.2 Facial Scores

### 4.2.1 Definition of Facial Scores

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

**Facial parts and facial-part sets.** Facial parts represent isolable facial components. We define facial part sets as $\mathcal{P} = \{P_1, ..., P_{N_p}\}$ 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 simple motions of facial parts. We define mode sets as $\mathcal{M}^{(a)} = \{M_1^{(a)}, ..., M_{N_a}^{(a)}\}$ where $N_a$ 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”.

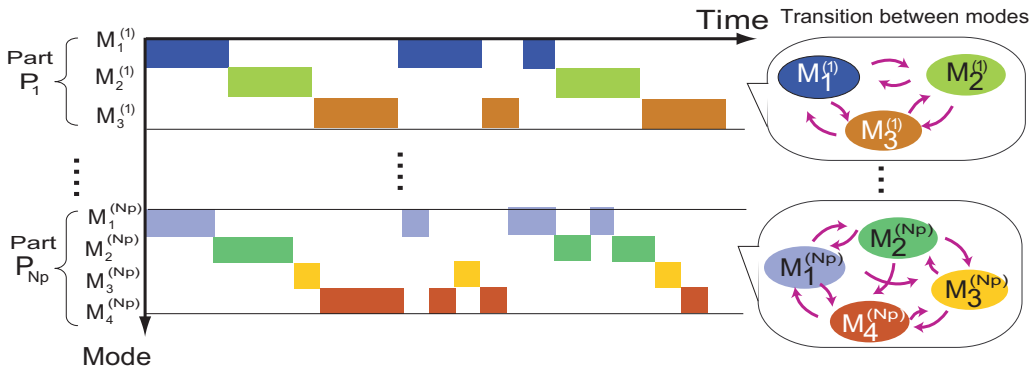



Figure 4.2: Facial scores. The vertical axis represents modes of facial parts, and the horizontal axis represents time. The transition of the motion of each facial part is described based on intervals along the temporal axis.

**Intervals and interval sets.** Intervals represent temporal ranges of modes. We define interval sets as $\mathcal{I}^{(a)} = \{I_1^{(a)}, ..., I_{K_a}^{(a)}\}$ where $K_a$ is the number of intervals into which time series data of a facial part $P_a$ is segmented. Intervals $I_k^{(a)} (k \in \{1, ..., K_a\})$ 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 length of time series data of a facial part $P_a$.

**Facial scores.** We define a facial score as a set that comprises all the interval sets of each facial parts $\{\mathcal{I}^{(1)}, ..., \mathcal{I}^{(N_p)}\}$. Figure 4.2 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. For each facial part, intervals with the same mode are depicted by the same color and aligned at the same row. Thus, the facial score describes timing structures among motions of the facial parts.

### 4.2.2 Facial Parts in Facial Scores

To recognize facial expression based on timing structures, we need to treat multiple facial areas where their movements are able to occur independently. Because the facial motion is produced by muscular action, a straight forward definition is to choose each muscle as a different part. However, some facial skin can be moved by multiple muscle action; moreover, some muscles are hard to control independently. We therefore use appearance-based definition.

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 [EF75]. We basically follow the Ekman's definition; that is, we use the three areas; in addition, we treat areas around the eyebrows and eyes on the left and right as different facial parts. This is 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. Here, we 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 $2n_a$-dimensional column vector:

$$x^{(a)} = (z_{x_1}^{(a)},\ z_{y_1}^{(a)}, ..., z_{x_{n_a}}^{(a)},\ z_{y_{n_a}}^{(a)})^\top, \tag{4.1}$$

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

### 4.2.3 Modes in Facial Scores

As we defined in Subsection 4.2.1, each complex movement of a facial part is composed of simple motion categories, which we refer to as 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 motions that have stable dynamics 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 Subsection 4.1.1. That is, all the modes are extracted by the clustering of dynamics from captured real data.

For a generative model of simple dynamics in each mode, we exploit the interval systems introduced in Chapter 2. The dynamics of the mode $M_i^{(a)} (i \in \{1, ..., N_a\})$ in a facial part $P_a$ is therefore modeled by the following linear dynamical system:

$$x_t^{(a)} = F^{(a,\,i)} x_{t-1}^{(a)} + g^{(a,\,i)} + \omega_t^{(a,\,i)}, \tag{4.2}$$

where $x_t^{(a)}$ is a internal state vector in a feature space at time $t$, $F^{(a,\,i)}$ is a transition matrix, which differs from other modes' matrices, $g^{(a,\,i)}$ 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, complex motion in each facial part is described based on the transition of linear dynamical systems, such as a hybrid dynamical system that we described in Chapter 2. Therefore, the proposed model can be considered as a concurrent process of multiple hybrid dynamical systems, where each hybrid dynamical system is applied to model dynamics in each part.

The extraction of mode is based on the clustering technique that we proposed in Section 3.3 in the previous chapter. We will briefly review the method here. Given a sequence of feature vectors, we first find a initial segmentation based on the velocity. We then merge the nearest dynamical system pairs iteratively based on agglomerative hierarchical clustering. A linear dynamical system, in general, can generate not only stable motions, which start from an initial shape and converge to a specific shape, but cyclic or oscillating motions. To extract only the stable motions, we proposed a method to provide a constraint on eigenvalues of the transition matrices.

## 4.3 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 relations as timing structures of the face. In this section, we describe a method to represent and extract timing structures from a facial score.

### 4.3.1 Definition of Timing Structures

We first concentrate on modeling timing structure between two parts $a$ and $b$. Let $I_{(i)}$ be an interval $I_k$ that has mode $M_i \in \mathcal{M}$ in part $P_a$ (i.e.,$m_k = M_i$), and let $b_{(i)}, e_{(i)}$ be its beginning and ending time points, respectively. (We omit index $k$, which denotes the order of the interval.) Similarly, let $I'_{(p)}$ be an interval that has mode $M'_p \in \mathcal{M}'$ in the range $[b'_{(p)}, e'_{(p)}]$ of part $P_b$. The temporal relation of two modes becomes the quaternary relation of the four temporal points $R(b_{(i)}, e_{(i)}, b'_{(p)}, e'_{(p)})$.

Here we break up the quaternary relation $R(b_{(i)}, e_{(i)}, b'_{(p)}, e'_{(p)})$ into the following four binary relations:

$$R_{bb}(b_{(i)}, b'_{(p)}), \quad R_{be}(b_{(i)}, e'_{(p)}),$$
$$R_{eb}(e_{(i)}, b'_{(p)}), \quad R_{ee}(e_{(i)}, e'_{(p)}).$$

Let us define timing structure as the relation $R$ that can be determined by a combination of these four binary relations above with respect to all the mode pairs $(M_i, M'_p) \in \mathcal{M} \times \mathcal{M}'$.

Considering temporal ordering relations $R_<, R_=, R_>$, which are often used in temporal logic [All83, All84, PMB97, PB97, Mas98], for these binary relations, we get $3^4$ relations for $R$. Because of $b_{(i)} \leq e_{(i)}$ and $b'_{(p)} \leq e'_{(p)}$, it can be reduced to 13 relations as shown in Figure 4.3(a). 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; that is, it is often important to analyze two motions in different facial parts are synchronized or not. We therefore utilize not only temporal order of events but metric information (i.e., scales and degree of temporal differences) among beginning and ending times of multiple intervals.

To extend the 13 categories using metric information, we use the temporal difference of two time points as the relation $R_D$, which can be represented by

metric $D \in \mathbf{R}$. Using this metric relation, we can define the first-order timing structure such as

$$D_{bb} = b_{(i)} - b'_{(p)} \qquad D_{be} = b_{(i)} - e'_{(p)}$$
$$D_{eb} = e_{(i)} - b'_{(p)} \qquad D_{ee} = e_{(i)} - e'_{(p)}$$

for $R_{bb}$, $R_{be}$, $R_{eb}$, and $R_{ee}$, respectively.

We can also define the second-order timing structure as the combination of two relations above. For example, the relation of $R_{bb}$ and $R_{ee}$ is represented by a point $(D_{bb}, D_{ee}) \in \mathbf{R}^2$ in a two-dimensional space if we use temporal difference $b_{(i)} - b'_{(p)}$ and $e_{(i)} - e'_{(p)}$ for the metric of $R_{bb}$ and $R_{ee}$, respectively. Figure 4.3(b) shows the two-dimensional space, where the horizontal and vertical axes represent the difference between the beginning times and the difference between the ending times, respectively. The interval pairs in the figure denote the typical temporal relations in each area of the two-dimensional space.

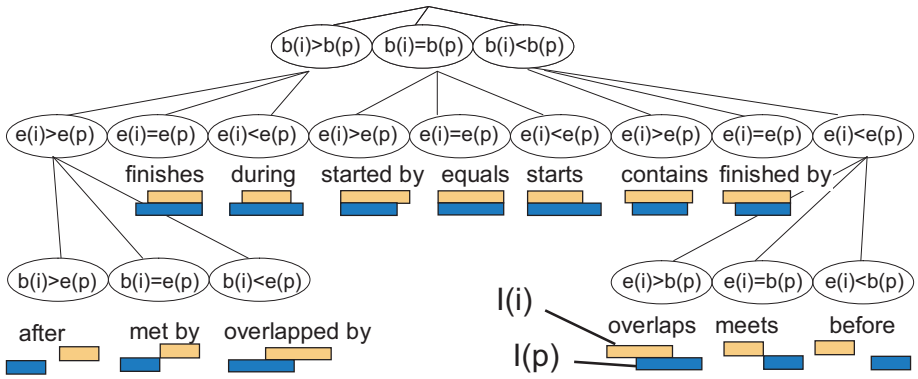
Note that, if we use the third-order timing structure that is defined by combination of three relations above, then all the temporal relations between the two intervals can be defined including the duration length.

### 4.3.2 Distributions of Timing Structures

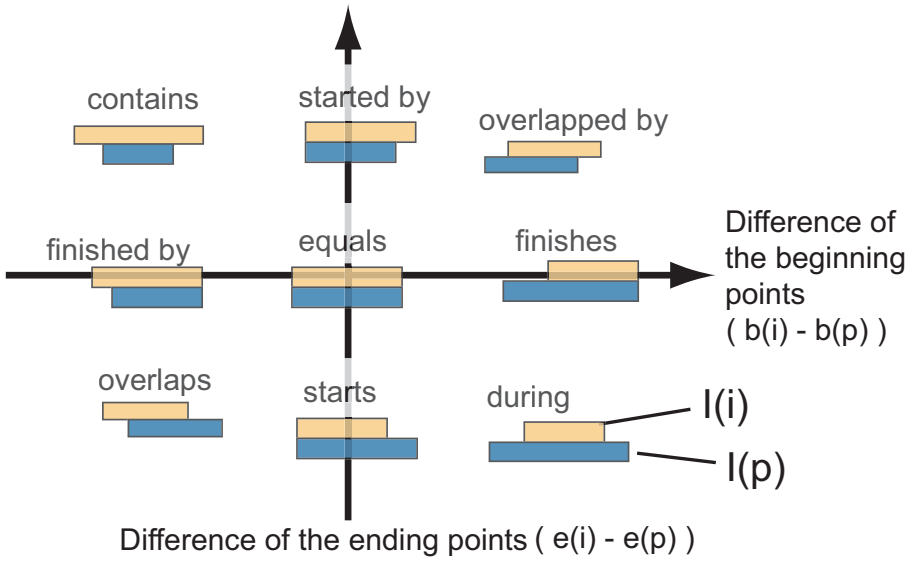
Using temporal differences between beginning and ending times, we can represent the distribution of first-order timing structure as four distributions $H(b_{(i)} - b_{(p)}), H(e_{(i)} - e_{(p)}), H(b_{(i)} - e_{(p)})$ and $H(e_{(i)} - b_{(p)})$, where $H(D)$ is a one-dimensional distribution of variable $D$. We can also represent the second-order timing structure as six two-dimensional distributions $H(b_{(i)} - b_{(p)}, e_{(i)} - e_{(p)}), H(b_{(i)} - b_{(p)}, b_{(i)} - e_{(p)}), H(b_{(i)} - b_{(p)}, e_{(i)} - b_{(p)}), H(e_{(i)} - e_{(p)}, b_{(i)} - e_{(p)}), H(e_{(i)} - e_{(p)}, e_{(i)} - b_{(p)})$ and $H(b_{(i)} - e_{(p)}, e_{(i)} - b_{(p)})$, where $H(D_1, D_2)$ is a two-dimensional distribution of variables $D_1, D_2 \in \mathbf{R}$. Representations of the third-order timing structures become three-dimensional distributions in the same manner.

## 4.4 Experiments

In this section, we evaluate the effectiveness of our representation by examining the separability between intentional smiles and spontaneous smiles using obtained facial scores from captured data.




(a) Temporal ordering relations.




(b) Temporal difference relation.

Figure 4.3: Temporal relations of two intervals. (a) The temporal order of beginning and ending time provides 13 relations of the two intervals. (b) The horizontal and vertical axes denote the difference between beginning points $b_{(i)} - b'_{(p)}$ and the difference between ending points $e_{(i)} - e'_{(p)}$, respectively.

### 4.4.1 Configuration of the Experiments

Intentional and spontaneous smiles of six subjects (we use ID A to F to distinguish them) were captured in 480 $\times$ 640 at 60 fps as the input image sequences. Then, we downsampled the images to 240 $\times$ 320 resolution. We used a camera system that was composed of a helmet and a camera fixed in front of the helmet to concentrate on the analysis of front faces. The camera system enabled us to capture front face images without self-occlusion even if large 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 during they were watching disgusting movie that have been standardized by Gross [GL95]. Spontaneous smiles were captured during they were watching Japanese-standup comedy (Manzai). Figure 4.4 (b) shows part of a captured face image sequence. The number of intentional smiles was 50 for all the subjects. The number of spontaneous smiles was different among the subjects: subject A, B, C, D, E, and F made 37, 39, 30, 38, 31, and 29 expressions, respectively.

### 4.4.2 Facial Feature Tracking

We tracked feature points in facial image sequences using the active appearance model (AAM) [CET98]. The AAM contains a statistical model of correlations between shape and grey-level appearance variation. The AAM-based feature point tracking consists of two stages. We first build an AAM model using a training set of face images and its feature points given manually. Then, we can use the model to extract facial feature points in novel images. Due to the trained model, AAM can search the feature points rapidly and robustly (see Appendix D for details).

Figure 4.4 shows an example of tracked feature points [1]. The number 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 jaw line (refer to Figure 4.4 (a)). Although the jaw line was not represented as one of the facial parts, it was used for improving tracking accuracy. Therefore, the dimensionality of feature vectors in the eyebrows (left/right), the eye (left/right), the nose, and the mouth were 10, 16, 22, and 16, respectively.

Figure 4.4 (c) shows part of a face image sequence with tracked feature points;

[1] Feature points were tracked using the AAM-API that Stegmann (Technical University of Denmark) developed [SG02].

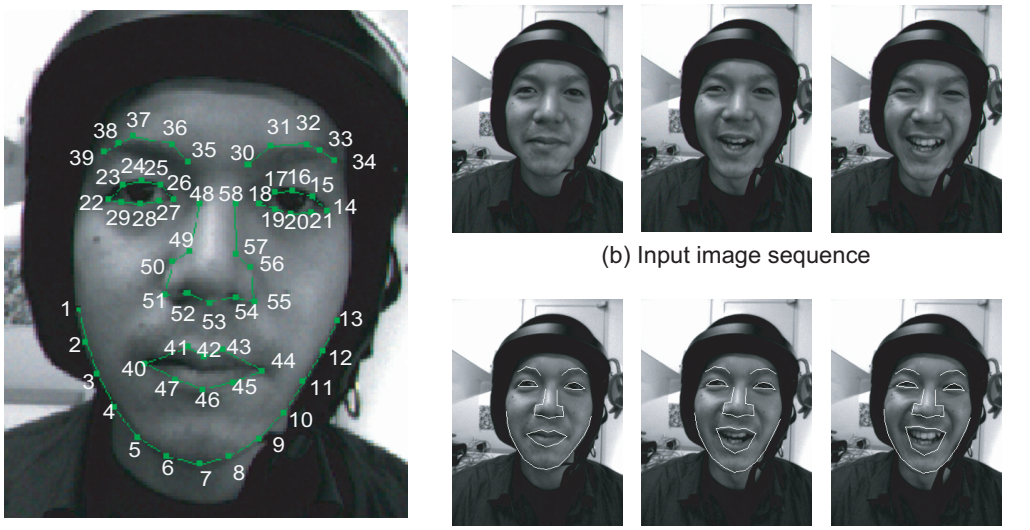



Figure 4.4: (a) A training image to build active appearance models. (b) Part of a captured face image sequence. (c) Part of a face image sequence with tracked feature points.

the frames correspond to the images shown in Figure 4.4 (b). Comparison of the corresponding images demonstrates precise detection of feature points in changes of facial expression.

### 4.4.3 Automatic Acquisition of Facial Scores

As we described in Subsection 4.2.3, the obtained feature vectors of each facial part were segmented into modes using the clustering of linear dynamical systems that we proposed in Section 3.3. Figure 4.5 is an example of the segmentation result of the mouth part. The vertical axes of the top, the middle and the bottom subfigures represent x-coordinates of feature points, y-coordinates of feature points and the transition of modes respectively. The horizontal axis of each subfigure represents time.

Figure 4.6 shows an example of the facial score that describes dynamic characteristics of all facial parts during intentional smiles. This figure suggests that the movement of each smile can be segmented into the following four modes: two stationary modes ("neutral" and "smiling") and two dynamic modes ("onset" and "offset" of smiling).

Figure 4.7 and Figure 4.8 shows the facial score of an intentional smile and a natural smile, respectively. We see that the beginning and ending timing of the in-

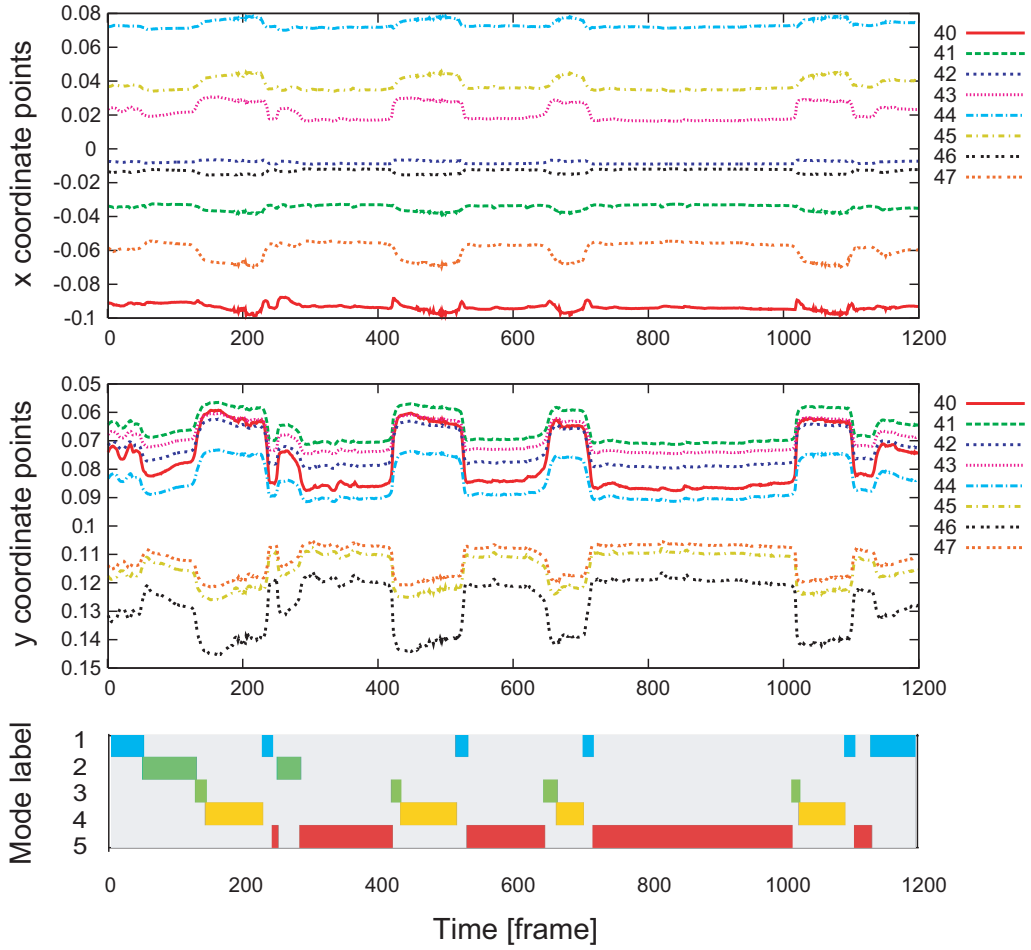



Figure 4.5: The correspondence of the mouth part of an obtained facial score from spontaneous smiles with the feature vector series. The vertical axes of the top, the middle and the bottom subfigures represent x-coordinates of feature points, y-coordinates of feature points and modes respectively, and 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 4.4 (a). For example, the mode 4 and 5 represent "remain open" and "remain closed", respectively.

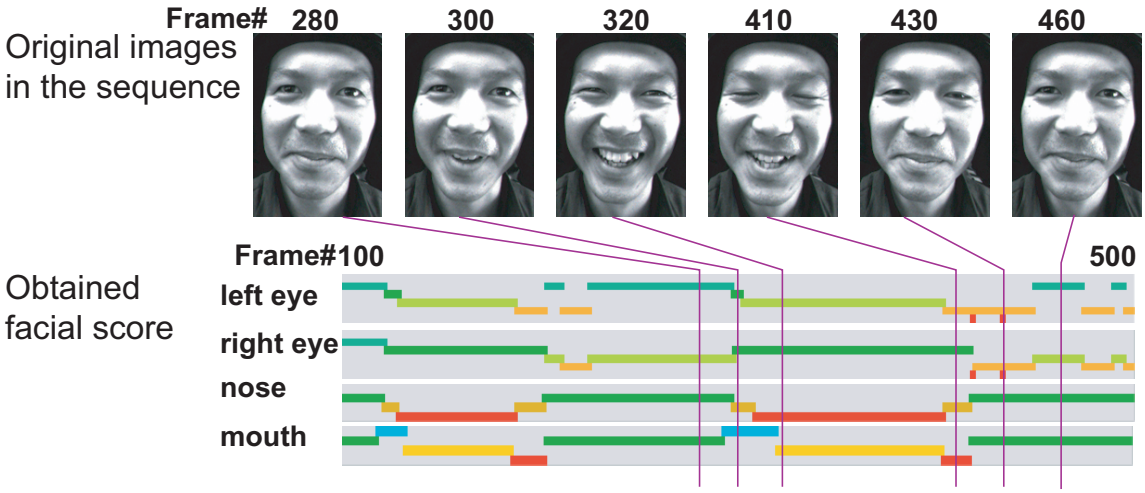



Figure 4.6: An example of obtained facial scores from intentional smiles (left and right eyebrows are omitted).

tervals are different in these expressions. Especially, the motions of the intentional smile are synchronized compared to the natural smile. In the next subsection, we evaluate the differences of these two smiles based on the comparison of timing structures.

Because of the limitation of the movie length and the capacity of the capturing system, we obtained facial expressions using several sessions. Then, we acquired facial scores from each of the sessions automatically. We however manually found the correspondence of modes among these sessions. In addition, we merged multiple intervals based on human observation in case that one mode was divided into multiple modes.

Despite that we could replace this manual operation with an automatic training method such as the expectation-maximization algorithm of the interval system described in Section 3.4, we chose to check and modify the segmentation results manually because we wanted to verify the effectiveness of timing structures rather than to evaluate the precision of the EM algorithm. To distinguish these two problems and to concentrate on verifying the effectiveness of timing structures, we postulated that the clustering algorithm provided a set of candidates for the segmentation, and we selected one of the candidates manually.

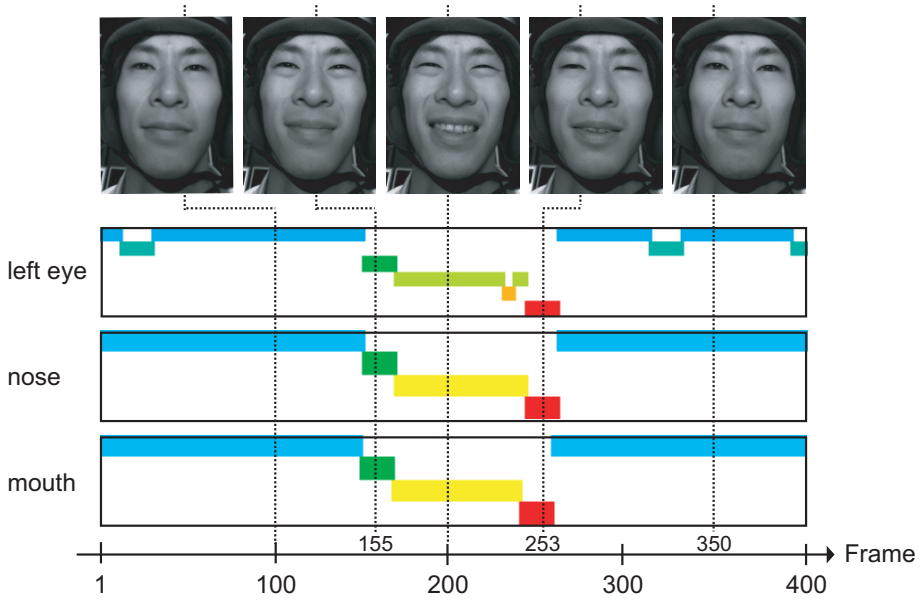



Figure 4.7: The facial score of an intentional smile.

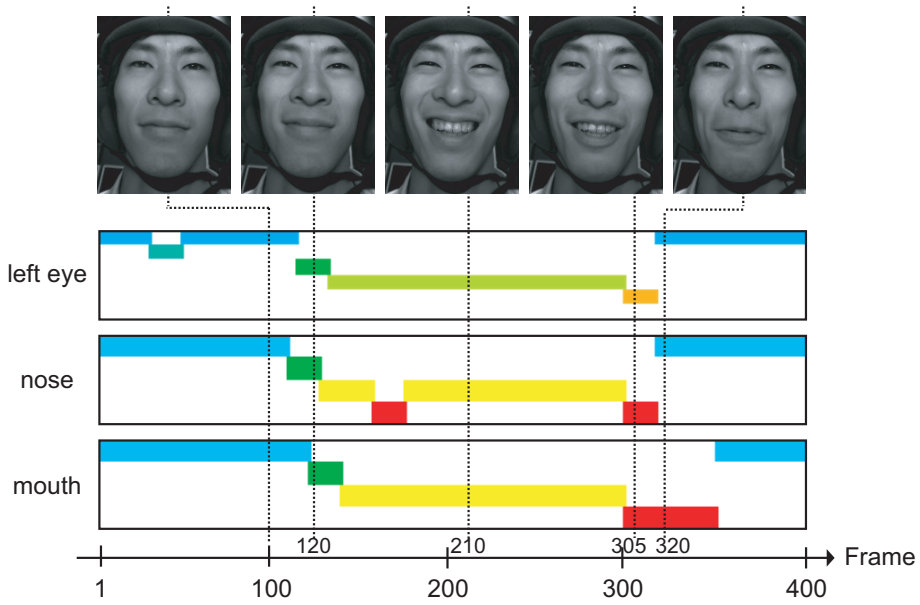



Figure 4.8: The facial score of a spontaneous smile.

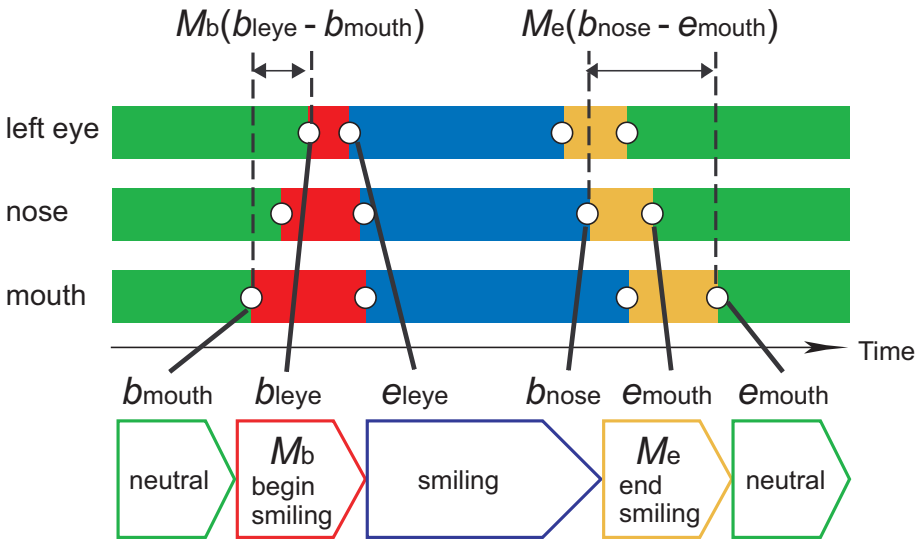



Figure 4.9: Onset mode $M_b$ and offset mode $M_e$.

### 4.4.4 Comparison of Timing Structures between Intentional and Spontaneous Smiles

#### Modes of Smile Onset and Offset

To evaluate the separability of intentional and spontaneous smiles using extracted facial scores, we defined the following two modes which is selected from automatically extracted modes:

$M_b$: onset motion of smiles (from neutral to smiling)

$M_e$: offset motion of smiles (from smiling to neutral)

To simplify the evaluation, we used a facial score that consists of three facial parts: left eye, nose, and mouth. In addition, since the duration lengths of stationary modes such as "neutral" and "smiling" closely depend on the context of the expression, we concentrate on analyzing timing structures among dynamic modes $M_b$ and $M_e$ (see Figure 4.9).

Let $b_{\text{leye}}$ and $e_{\text{leye}}$ be the beginning and ending time points of the left eye motion in its facial score. Similarly, let $b_{\text{nose}}$ and $e_{\text{nose}}$ be those of the nose motion, and $b_{\text{mouth}}$ and $e_{\text{mouth}}$ be those of the mouth motion, respectively. Then we extract temporal differences between such time points; for example, we use $M_b(b_{\text{nose}} - b_{\text{mouth}})$, which denote the temporal difference between the beginning of nose motion and that of mouth motion during the onset of a smile.

### Analysis of Timing Structures

To analyze the intentional and natural facial expression, we exploited the representation of timing structure described in Section 4.3. We first used the first-order timing structure analysis using one-dimensional distributions as preliminary experiments. Since this preliminary experiments showed that any single temporal difference cannot discriminate the two smile categories (i.e., intentional and spontaneous), we employed a pair of temporal differences (i.e., the second-order timing structure) as a distinguishing feature. That is, a feature to characterize each shot of smile is represented by a point in the two-dimensional space whose axes denote a selected pair of temporal differences. Since there are many possibilities for the combination of temporal differences, for each pair of temporal differences, we calculated the Maharanobis generalized distance between a pair of distributions of two smile categories, and selected such pair of temporal differences that the two distributions took the largest distance. Note that since smiling actions may differ from person to person, we extracted a distinguishing feature for each subject person.

Figure 4.10 shows the experimental results for six persons. Each subfigure shows the selected two-dimensional space and the distributions of intentional and natural smile categories for each subject. Each point denotes a single expression. From this figure, we observe that we can discriminate the distributions of two smile categories using their dynamic features.

To evaluate the effectiveness of timing structure for discriminating intentional and natural smiles, we calculated recognition rate of each smile for each subject based on leave-on-out method [DHS00]. First, we trained a linear discriminate boundary plane using support vector machines [Bur98]; we then discriminated the test data. The result is shown in Table 4.1. We see that all the recognition rates for all the subjects are in the ranges from 79.4% to 100%. Hence, the timing structure provides an enough feature for distinguishing intentional and natural smile categories.

### Differences among the Subjects

From Figure 4.10, we see that the extracted axes are different among the six subjects. Especially, the axes that correspond to duration lengths of onset or offset motions (e.g., $M_b(b_{\text{nose}} - e_{\text{mouth}})$) were extracted from five subjects excluding subject C.

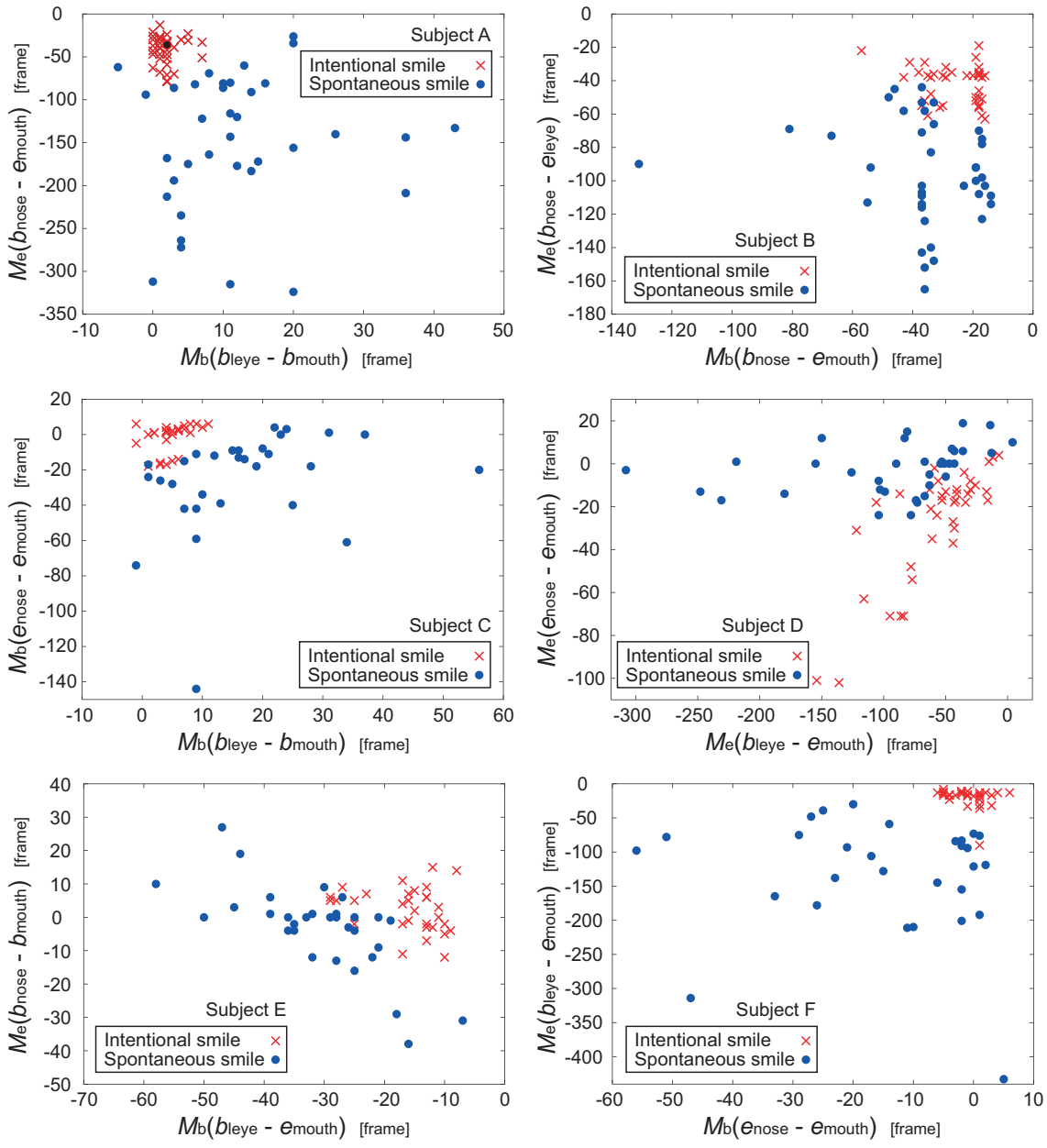



Figure 4.10: Timing structure with the longest distance between intentional and spontaneous smiles distribution.

Table 4.1: Accuracy of discrimination between intentional and spontaneous smiles based on timing structures in Figure 4.10.

| **Subject** | **Intentional** (%) | **Spontaneous** (%) |
|---|---|---|
| A | 100 | 83.8 |
| B | 100 | 79.4 |
| C | 82.4 | 96.4 |
| D | 85.1 | 79.7 |
| E | 85.3 | 90.3 |
| F | 96.6 | 93.1 |

For subject A and C, the axis that denotes the difference of the beginning points between the left eye and mouth parts in the onset motion (i.e., $M_b(b_{\text{leye}} - b_{\text{mouth}})$) were extracted. This feature corresponds to the feature that is used in the psychological experiments conducted by Nishio, et al. [NKN98]. Therefore, intentional smiles of subject A and C might be easily discriminated from natural smiles by human.

Consequently, the personality of facial expression becomes significant especially for discriminating intentional and spontaneous expression compared to emotion recognition in which most of the existing work attempts to find common factor of facial expression.

## 4.5 Discussion

In this chapter, 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 each of which is characterized by linear dynamical system, and the effectiveness of the timing structure for discriminating the two smile categories.

To emphasize the characteristics of the proposed representation, we focused on using only timing structures. However, other features of movement such as scale, speed and duration, which provide further information on recognizing fa-

cial 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 produced by personal habits. Directions for future work are to tackle these problems and to evaluate the effectiveness of timing structure using a large number of captured sequences.