Recognizing Communicative Facial Expressions for Discovering Interpersonal Emotions in Group Meetings

Shiro Kumano
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya, Atsugi-shi
Kanagawa, Japan
kumano@eye.brl.ntt.co.jp

Kazuhiro Otsuka
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya, Atsugi-shi
Kanagawa, Japan
otsuka@eye.brl.ntt.co.jp

Dan Mikami
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya, Atsugi-shi
Kanagawa, Japan
mikami.dan@lab.ntt.co.jp

Junji Yamato
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya, Atsugi-shi
Kanagawa, Japan
yamato@brl.ntt.co.jp

ABSTRACT

This paper proposes a novel facial expression recognizer and describes its application to group meeting analysis. Our goal is to automatically discover the interpersonal emotions that evolve over time in meetings, e.g. how each person feels about the others, or who affectively influences the others the most. As the emotion cue, we focus on facial expression, more specifically smile, and aim to recognize "who is smiling at whom, when, and how often", since frequently smiling carries affective messages that are strongly directed to the person being looked at; this point of view is our novelty. To detect such communicative smiles, we propose a new algorithm that jointly estimates facial pose and expression in the framework of the particle filter. The main feature is its automatic selection of interest points that can robustly capture small changes in expression even in the presence of large head rotations. Based on the recognized facial expressions and their directions to others, which are indicated by the estimated head poses, we visualize interpersonal smile events as a graph structure, we call it the interpersonal emotional network; it is intended to indicate the emotional relationships among meeting participants. A four-person meeting captured by an omnidirectional video system is used to confirm the effectiveness of the proposed method and the potential of our approach for deep understanding of human relationships developed through communications.

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

Categories and Subject Descriptors

H1.2 [Models and Principles]: User/Machine System—Human Information Processing

General Terms

ALGORITHMS, HUMAN FACTORS

Keywords

meeting analysis, facial expression, direction of facial expression, interpersonal emotion

1. INTRODUCTION

Face-to-face conversation is one of the most basic forms of communication in daily life for sharing information, understanding others' emotion, and making decisions etc. In the face-to-face setting, people exchange not only verbal messages but also nonverbal messages using multimodal channels such as prosody, gaze, body posture, and facial, head, and hand gestures; the importance of these exchanges has been psychologically elucidated [1]. In recent years, multimodal meeting analysis has been acknowledged as an emerging research area and intensive efforts have been made to analyze meetings [7]; the existing studies mainly focused on recognizing relatively low-level visible and auditory behaviors such as speaker diarization, tracking face position and pose, gaze directions. Although a few studies have attempted to understand higher-level meeting states, e.g. conversation regimes such as monologue or dialogue [14], and the dominant person in meeting [10], the emotional aspect of meetings has hardly been addressed in the field of automatic meeting analysis.

Among emotional expressions in meetings, smile is expected to be quite important, because it plays key roles in regulating conversation flow, expressing positive feeling, and building and maintaining intimacy or rapport [3]. Smile mainly differs from laughter, e.g. tackled in [17], in its lack

of vocalization and more subtle exposure [23]. This indicates smile can easily coexist with the speaker's utterance and others' smile. More notably, some psychological works indicate that smile has another significant characteristic in conversations; smile is frequently directed to someone, that is, a spontaneous smile is a reliable sign of positive feelings towards a specific receiver [18], while laughter seems to be only weakly directed. Some studies have addressed the automatic detection of spontaneous smiles, exposed in non-meeting situations [4, 9, 22]. Especially in [4, 22], head movement information is additionally utilized for improving the accuracy of smile detection. Other studies targeted the detection of laughter [17] and interest [19] in meetings. However, no paper has addressed the direction of facial expression, namely the intended recipient.

Recognizing communicative smiles in meetings from visual sources, or images, is not easy, because smiles generally involve the subtle motion of facial parts and audio cues are basically absent. Furthermore, although participant's head movements to look at other participants are important cues for assigning the direction of facial expressions, they make vision-based facial expression recognition (FER) more difficult. Large head rotation yields slanted face views, and such distortion hampers the stable recognition of subtle expressions. Note that assuming near-frontal-view faces, as is done by most existing FER methods found in some excellent reviews [6, 21, 16, 25], is impractical in the meeting situation, because participants often turn their faces to look at other participants. Accordingly, facial expression cues should be extracted in the presence of head movements, as mentioned in [5, 11].

In addition, the interpersonal difference of spontaneous smiles should also be handled to correctly discriminate smile from laughter or other visually-similar expressions. The popular approach in vision-based FER is to prepare a single general model for facial expression and apply it to arbitrary users [2, 5]. However, it's well reported in [8] that the overfitting problem makes it difficult to create a completely accurate general model. To avoid it, the method proposed in [11] utilizes a simple person-specific face model, called the variable-intensity template, for simultaneously estimating facial pose and expression. The variable-intensity template describes how the intensities of multiple points, defined in the vicinity of facial parts, vary with different facial expressions. However, interest points that can well discriminate target expressions may not be selected in their method, hence smiles are expected to be often confused with laughter.

Against this background, as the first step in an unexplored research topic in automatic meeting analysis, i.e. discovering interpersonal emotions, this paper addresses the recognition of smile events evolved over time in meetings, or "how often who is smiling at whom". To this end, we first propose a novel joint estimation method of facial expression and head pose that can discriminate spontaneous smiles from laughter and other facial expressions from images. Their directions are then inferred from the estimated head poses, based on the experimental finding in [20] that head orientation is a sufficient indicator of the participant' focus of attention. Finally, by using the recognized facial expressions and their directions to others, we visualize interpersonal smile events as a graph structure, which we call the interpersonal emotional network. The interpersonal emotional network is intended to indicate the emotional relationships among meeting participants.

Although the FER part of the proposed method is mainly based on [11], the key mechanism of interest point selection for discriminating visually similar but functionally different facial expressions is newly proposed in this paper; the points that are most indicative of each target facial expression against head pose variations are selected. The proposed method can recognize spontaneous but subtle facial expressions in meetings, where the participants frequently and significantly moved their heads toward others.

A four-person round-table meeting captured by an omnidirectional video system is used to confirm the effectiveness of the proposed method and the potential of our approach to provide deep understanding of human relationships developed through communications.

The remainder of this paper is organized as follows. First, Section 2 describes our proposed method. Next, in Section 3, experimental results are given. Finally, a summary and future work are given in Section 4.

2. PROPOSED METHOD

The proposed method consists of training and inference stages, as shown in Fig.1. In the training stage, a face model, an extended version of the variable-intensity template in [11], is created for each person from training video sequence labeled with target facial expression categories. In the inference stage, first, head pose and facial expression are simultaneously estimated. The direction of the facial expression is then estimated based on the head pose. All of these processes are performed individually for each participant. Finally, an interpersonal emotion network, namely a network describing 'how often who is smiling at whom?', is created from the estimated expressions and their directions of all participants.

Main symbols used afterwards are listed here. Facial expression and its direction are represented by e_i and g_i, where i denotes participant $i \in \{1, \cdots, N_i\}$. The facial expression e denotes a category, i.e. $e \in \{1, \cdots, N_e\}$. The state of the direction of facial expression is also quantized, i.e. $g \in \{1, \cdots, N_i\}$. Direction $g_{i,t} = j(\neq i)$ means the facial expression of P_i at time t is directed to P_j, while $g_{i,t} = i$ means the direction is averted, i.e. towards none of the other participants. This paper sets the number of participants, N_i, and the number of target expressions, N_e, at four and three respectively in this paper: $e \in$ {smile, laughter, others}. In short, the state of $e_{i,t}$ = 'smile' and $g_{i,t} = j(\neq i)$ means that participant P_i smiled at participant P_j at time t.

2.1 Problem formulation for joint estimation of facial pose and expression

In our framework, head pose and facial expression are estimated by calculating their likelihood given face images. Figure 2 describes their relationship in our model, where head pose $\boldsymbol{h}_t$ and facial expression e_t are assumed to be conditionally dependent given image $\boldsymbol{z}_t$. Furthermore, head pose and facial expression are assumed to follow individual Markov processes. Head pose $\boldsymbol{h}$ has 6-DOF, i.e. the position in the image (2D) and scale of the face, and three-dimensional orientation angles (yaw, pitch, and roll).

Estimators of head pose and facial expression at time t, $\widehat{\boldsymbol{h}}_t$ and $\widehat{e}_t$, are calculated based on their joint posterior probability density function (pdf) at time t, $p(\boldsymbol{h}_t, e_t|\boldsymbol{z}_{1:t})$. The

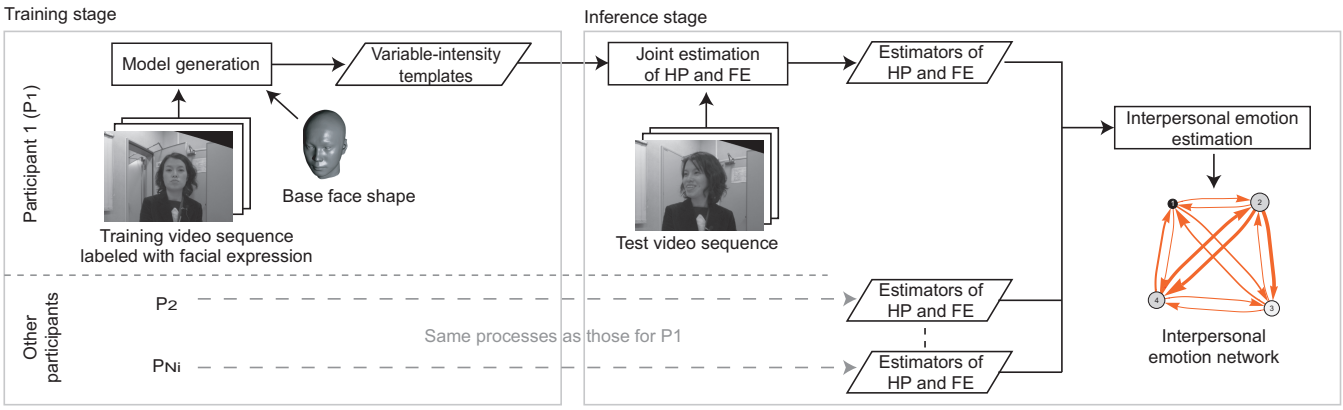


Figure 1: System flow chart: The proposed method consists of training and inference stages. In the training stage, a variable-intensity template for each person is created from training video sequence labeled with target facial expression categories. In the inference stage, first, head pose and facial expression are simultaneously estimated individually for each person by utilizing his/her variable-intensity template. Then, interpersonal emotion is inferred from the estimated head poses and expressions of all participants. Moreover, HP and FE in the figure denote head pose and facial expression, respectively.

head pose estimator is defined to be the expectation of its marginal posterior pdf. The recognized expression is obtained as the expression that maximizes its marginal posterior probability mass function (pmf):

$$\widehat{\boldsymbol{h}}_t = \mathrm{E}_{p(\boldsymbol{h}_t|\boldsymbol{z}_{1:t})}[\boldsymbol{h}_t] \quad (1)$$

$$\widehat{e}_t = \arg\max_{e_t} P(e_t|\boldsymbol{z}_{1:t}). \quad (2)$$

The joint posterior pdf is decomposed by the following Bayes' rule:

$$p(\boldsymbol{h}_t, e_t|\boldsymbol{z}_{1:t}) \propto p(\boldsymbol{z}_t|\boldsymbol{h}_t, e_t)p(\boldsymbol{h}_t, e_t|\boldsymbol{z}_{1:t-1}) \quad (3)$$

where $p(\boldsymbol{z}|\boldsymbol{h}, e)$ denotes the joint likelihood function of head pose $\boldsymbol{h}$ and facial expression e for input face image $\boldsymbol{z}$, and $p(\boldsymbol{h}_t, e_t|\boldsymbol{z}_{1:t-1})$ represents the predictive distribution at time t. The joint likelihood is defined as the product of the likelihood of the head pose for the face image and that of the facial expression for the head pose and face image in the proposed method: $p(\boldsymbol{z}|\boldsymbol{h}, e) = p(\boldsymbol{z}|\boldsymbol{h})p(\boldsymbol{z}, \boldsymbol{h}|e)$. These likelihoods are defined based on our variable-intensity template. Details of the likelihood are given in 2.2. Moreover, the posterior pdf is normalized as $\int \sum_{e_t} p(\boldsymbol{h}_t, e_t|\boldsymbol{z}_{1:t})\mathrm{d}\boldsymbol{h}_t = 1$.

Conditioning the predictive distribution on the head pose and the expression yields the following recursive form:

$$p(\boldsymbol{h}_t, e_t|\boldsymbol{z}_{1:t-1}) = \int p(\boldsymbol{h}_t|\boldsymbol{h}_{t-1}) \sum_{e_{t-1}} P(e_t|e_{t-1}) \, p(\boldsymbol{h}_{t-1}, e_{t-1}|\boldsymbol{z}_{1:t-1})\mathrm{d}\boldsymbol{h}_{t-1} \quad (4)$$

where $p(\boldsymbol{h}_t|\boldsymbol{z}_{1:t})$ represents the posterior pdf of head pose given all face images up to time t. Transition matrix between expressions, $P(e_t|e_{t-1})$, is currently set to be equal for all expression combinations, i.e. no prior knowledge of facial expression transitions is assumed, in this paper. However, any expression transition matrix can be utilized in the proposed framework. On the other hand, the dynamics of the head pose is expressed by the memory of poses [12].

The predictive distribution in Eq.(4), unfortunately, cannot be calculated exactly in a closed-form, because of the nonlinearity of the camera projection function, partial occlusion of the face etc. Accordingly, this paper adopts the particle filter, or the so called sequential Monte Carlo method, to obtain an approximation of the posterior pdf.

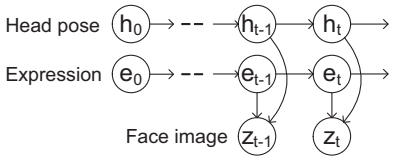


Figure 2: Dynamic Bayesian Network describing the relationship between head pose, facial expression, and face image.

2.2 Variable-intensity template

Likelihood $p(\boldsymbol{z}|\boldsymbol{h}, e)$ is defined based on our variable-intensity template $\mathcal{M}$, which consists of the following three kinds of components: $\mathcal{M} = \{\mathcal{S}, \mathcal{P}^{\mathrm{E}}, \mathcal{P}^{\mathrm{H}}, \mathcal{I}^{\mathrm{E}}, \mathcal{I}^{\mathrm{H}}\}$ where $\mathcal{S}$, $\mathcal{P}$, and $\mathcal{I}$ denote a rigid face shape model, a set of interest points, and an intensity distribution model, respectively. Superscripts E and H indicate facial expression and head pose, respectively.

The set of interest points $\mathcal{P}^*$, where the target $* \in \{\mathrm{E}, \mathrm{H}\}$, describes the position of focus on the face in calculating the target likelihood. These interest points are sparsely defined in the training images. In particular, the set of interest points for facial expression recognition, $\mathcal{P}^{\mathrm{E}}$, are defined at the locations where the intensity is salient for a specific facial expression. The set of interest points is described as: $\mathcal{P}^* = \{\boldsymbol{p}_k^*\}_{k=1}^{N_k^*}$, where $\boldsymbol{p}_k$ denotes the image coordinates of the k-th interest point in the training images, described later. N_k^* denotes the number of interest points for the target $*$. On the other hand, the intensity distribution model $\mathcal{I}^*$ describes the intensity distribution of each interest point k for different facial expressions: $\mathcal{I}^* = \{\mathcal{I}_k^*\}_{k=1}^{N_k^*}$. The interest points $\mathcal{P}^*$ and their intensity distribution models $\mathcal{I}^*$ are person-specific, that is, they are created for each person from his/her own training images. This enables us to

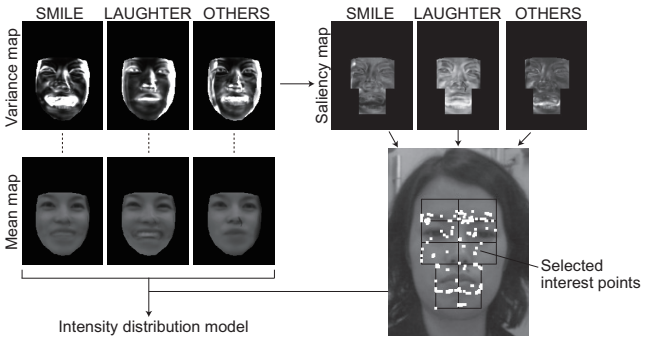


Figure 3: Flow of selection of interest points and generation of intensity distribution model for facial expression: Interest points having high saliency are selected one-by-one in each facial part. Note that only a few points are selected at the edges of facial parts, where the intensity varies significantly due to head tracking error and variation in shift of the facial part.

discriminate subtle facial expressions regardless of the large interpersonal variations.

Face shape model $\mathcal{S}$ gives the depth-information of each interest point for projecting it onto input face images according to head pose. Face shape model $\mathcal{S}$ is created by stretching a general basic face shape to fit the face in the training image Y_0 [12]. As the basic shape, the average face shape model[1], shown in Fig.1, is used in this paper.

The training image set consists of multiple images labeled with facial expression category: $\boldsymbol{Y} = \{Y_0, \boldsymbol{Y}_{e=1}, \ldots, \boldsymbol{Y}_{e=N_e}\}$, where Y_0 denotes single image in neutral expression, and $\boldsymbol{Y}_e$ denotes a set of multiple images labeled with facial expression $e(> 0)$. Image Y_0 is used for training in head pose tracking and images $\boldsymbol{Y}_e$ are used for facial expression recognition. The face in each image is frontal and is aligned between images. How to generate such training images is described in 2.2.3.

2.2.1 Models for facial expression recognition

The set of interest points for facial expression recognition, $\mathcal{P}^{\mathrm{E}}$, is defined at the locations where the intensity is salient for a specific facial expression. They are selected one by one in the four facial part regions (eyebrows, eyes, nose, and mouth) in training image Y_0, where the intensity is salient for one of the target facial expressions. The number of interest points for each expression is set to be 68 (eyebrows: 12×2, eyes: 6×2, nose: 8, and mouth: 24). That is, the total number of interest points, N_k^{E}, is 204. This number was decided after conducting preliminary evaluations of the performance metrics of accuracy and processing speed. Figure 3 shows an example of the selected interest points. Note that different subjects are likely to demonstrate different locations of the interest points due to the difference in face appearance and the movement of each facial part. Moreover, these facial part regions are roughly detected to be rectangular boundaries by using a cascaded AdaBoost detector based on Haar-like features [24].

The intensity distribution model for FER, $\mathcal{I}^{\mathrm{E}}$, describes how the interest point intensity varies for different facial expressions. Focusing on this property, we recognize facial expressions from the changes in observed interest point intensities. The observed intensity of each interest point varies due to localization error of the interest point caused by error in the shape model, change in intensity due to head orientation variation etc. The variation of each location on the face is represented as a normal distribution in our method. That is, the intensity distribution model of the k-th interest point, $\mathcal{I}_k^{\mathrm{E}}$, is described as:

$$\mathcal{I}_k^{\mathrm{E}} = \mathcal{N}\left(\mu_k(e), \sigma_k^2(e)\right) \tag{5}$$

where $\mu_k(e)$ and $\sigma_k^2(e)$ denote the mean and variance of intensity of the k-th interest point for facial expression e, respectively. The mean and variance are set to be those for the training image set $\boldsymbol{Y}_e$, shown as mean and variance maps in Fig.3, at location $\boldsymbol{p}_k^{\mathrm{E}}$.

Interest point selection.

Saliency of location $\boldsymbol{x}$, or coordinates in face image, for facial expression e is defined as the ratio of the variance between facial expression categories to the variance within category e:

$$S_{\boldsymbol{x},e} = \sigma^2_{\boldsymbol{x},B(e)} / \sigma^2_{\boldsymbol{x},W(e)} \tag{6}$$

where $\sigma^2_{\boldsymbol{x},W(e)}$ denotes the variance of intensity at position $\boldsymbol{x}$ within classes, where category e is assigned to one class and other categories are grouped in the other class. Variance $\sigma^2_{\boldsymbol{x},B(e)}$ is the variance between these two classes. These are calculated based on the variance for each expression category, i.e. the value in the variance map for each expression, shown in Fig.3, at position $\boldsymbol{x}$.

The interest point selection for facial expression follows the next procedure: First, saliency for the target expression e at all locations $\boldsymbol{x}$, $S_{\boldsymbol{x},e}$, is calculated from the training image set for expression e, $\boldsymbol{Y}_e$. Second, the point with the largest $S_{\boldsymbol{x},e}$ is picked up as a new interest point. Third, the saliency of points neighboring the newly added interest point are decreased according to distance, d, from the added point to the target pixel: $S_{\boldsymbol{x},e} \leftarrow S_{\boldsymbol{x},e} - \alpha \cdot \exp\left(-d^2\right)$. If the total number of selected pairs does not reach the limit, the selection process is reentered at the second step.

Likelihood of facial expression.

The likelihood of facial expression e for head pose $\boldsymbol{h}$ and face image $\boldsymbol{z}$, $p(\boldsymbol{h}, \boldsymbol{z}|e_t)$, is defined based on the variable-intensity template.

Assuming that the intensities of the interest points are independent, the likelihood is decomposed into the likelihood for each interest point:

$$p(\boldsymbol{h}, \boldsymbol{z}|e) = \prod_{k \in \mathcal{P}^{\mathrm{E}}} p(z_k(\boldsymbol{h})|e) \tag{7}$$

where $z_k(\boldsymbol{h})$ denotes the intensity at the location of the k-th interest point under head pose $\boldsymbol{h}$ in the face image.

$$p(z_k|e) = \frac{1}{\sqrt{2\pi}\sigma_k(e)} \exp\left[-\frac{1}{2}\rho\left(d_k\right)\right], \tag{8}$$

$$d_k = \frac{z_k - \mu_k(e_t)}{\sigma_k(e_t)} \tag{9}$$

[1]Average head dummy of Japanese young males contains over 100,000 polygons without texture. It is published by Digital Human Research Center, Advanced Industrial Science and Technology, http://www.dh.aist.go.jp/research/centered/facedummy/.

where function $\rho(\cdot)$ denotes a robust function. In this paper, we use the Geman McClure function with scaling factor $c(=9)$ which regulates an infinite input: $\rho(x) = c \cdot x^2/(1+x^2)$. This robust function makes the estimation more proof against noise such as imaging noise, and large position shifts due to shape model error.

Intensity $z_{i,t}$ is obtained as the intensity of face image $\boldsymbol{z}_t$ at the coordinate of the k-th interest point under head pose $\boldsymbol{h}_t$. The image coordinate is obtained via a three-step process: (1) orthogonal projection from the training image plane onto the shape model $\mathcal{S}$, (2) translation and rotation of $\mathcal{S}$ according to pose $\boldsymbol{h}_t$, and (3) projection of interest point i on shape model $\mathcal{S}$ onto the target image plane.

2.2.2 Models and likelihood for head tracking

It is hard to obtain an intensity distribution of each position of the face before tracking, because the location of the target point in each image moves with the head pose. Accordingly, the likelihood of head pose is defined as the difference in intensity of interest points $\mathcal{P}^{\mathrm{H}}$ between the input image and the training image in neutral expression Y_0. These components for head tracking are mainly based on a similar existing head tracker [12]. The tracker can successfully recover from tracking lost caused by the quick head motion, generated when changing the visual focus of attention from one participant to another, or self-occlusion of the face by hand gestures etc. These actions will occur frequently in meetings.

Interest points for head tracking, $\mathcal{P}^{\mathrm{H}}$, are sparsely selected as dipoles straddling the edges on the face, where the difference in intensity across the dipole is large. The number of interest points for head pose tracking, N_k^{H}, is set to be 256 in this paper. The resulting interest points are widely distributed all over the face, unlike the interest points for facial expression recognition shown in Fig.3. For the intensity distribution model, $\mathcal{I}^{\mathrm{H}}$, the mean $\mu_{k,0}$ is set to be the intensity of the neutral face image at location $\boldsymbol{p}_k$, and the standard deviation is assumed to equal the mean, $\sigma_{k,0} = \mu_{k,0}$.

The head pose likelihood is given as [12]:

$$p(\boldsymbol{z}|\boldsymbol{h}) = 1/\sum_{k\in\mathcal{P}^{\mathrm{H}}} \rho\left(d_k\right) \tag{10}$$

where scaling factor c in robust function ρ is set to be one. Moreover, in the head tracking process, input intensity z is adjusted globally all over the face to cancel the intensity change due to a change in head orientation.

2.2.3 Training images

The training images are created as follows: First, the variable-intensity template except for facial expression, or $\{\mathcal{S}, \mathcal{P}^{\mathrm{H}}, \mathcal{I}^{\mathrm{H}}\}$ is created for each participant from a training video sequence. The head of each participant is then tracked for all frames in the video sequence by utilizing the template. This tracking procedure is based on that in [12]. Next, the frontal face images are created by projecting the sampling points, or pixel, of the frontal face backward onto the input face images according to head pose. Note that, the size of the frontal face is set to be 150×200 in this paper.

2.3 Estimation of interpersonal emotions

Interpersonal emotion is estimated from which facial expressions are made and for how long. The following indicators are defined mainly with reference to the work in [15], where interpersonal influence in conversations were quantified without handling facial expressions.

2.3.1 Direction of facial expression

We quantize direction of facial expression g in the maximum likelihood scheme, by assuming each person's head pose follows a normal distribution given the target person to whom he/she exposed the facial expression:

$$\widehat{g}_i = \arg\max_j \mathcal{N}(h_i^{\mathrm{HOR}}; \kappa \cdot \phi_{i,j}, \sigma^2) \tag{11}$$

where $\widehat{g}_i = j$ means the estimated direction of facial expression of person P_i is looking at other person $P_j (j \neq i)$. Angle h_i^{HOR} denotes horizontal head orientation of person P_i, $\phi_{i,j}$ denotes the relative face angle from person P_i to P_j, and σ^2 and κ are a variance and a scaling factor, respectively. In addition, the likelihood function representing the person averting his/her gaze from everyone is defined as a uniform distribution. Note that the relative face angle between each participant pair in an omnidirectional image of round-table meeting, described in Section 3, can be simply calculated from their estimated face position and its orientation in the image, by assuming that the distance to each person from the camera is the same [13].

2.3.2 Interpersonal emotion network

The interpersonal emotion network is defined using the amount of smiling of person P_i directs to person P_j during a conversation. The amount of smiling of P_i is defined as the normalized duration of the smiling of P_i while gazing at P_j:

$$S_{i\to j} = \#\{\widehat{e}_{i,t} = \text{smile}, \widehat{g}_{i,t} = j\}_{t=1}^{T}/T \tag{12}$$

where $\#\{\cdot\}$ denotes the number of times wherein the conditions in the brackets are fulfilled, and T is the total number of video frames. Large $S_{i\to j}$ suggests that person P_i tries to attract the attention of person P_j.

2.3.3 Characteristics of each participant

Total amount of smiling that person i receives from all others, $S_{\mathrm{IN}}(i)$, and that person i gave to others, $S_{\mathrm{OUT}}(i)$, are defined as:

$$S_{\mathrm{IN}}(i) = \sum_{j(\neq i)} S_{j\to i}, \qquad S_{\mathrm{OUT}}(i) = \sum_{j(\neq i)} S_{i\to j}. \tag{13}$$

Large $S_{\mathrm{OUT}}(i)$ with a small variance between all other participants suggests that P_i tried to globally regulate the conversation flow or just to be well-liked by everyone. Large $S_{\mathrm{OUT}}(i)$ with a large variance indicates that P_i likes only some participants. Small $S_{\mathrm{OUT}}(i)$ indicates that person P_i is not active in regulating the conversation flow or is not interested in the conversation.

3. EXPERIMENTAL RESULTS

This paper targets four-person group conversations. The participants were four women within the same age bracket, seated as shown in Fig.4. They were instructed to hold a discussion and try to reach a conclusion as a group on a given discussion topic within eight minutes. The discussion topic was "Should smoking be fully prohibited in public spaces?". The conversation was captured with a tabletop sensing device for roundtable meetings [13]; it consists of

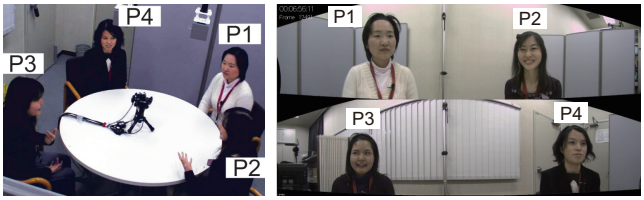


Figure 4: Example scene of the conversation. Left: Overall view taken by an extra handy camera. Right: Omnidirectional view captured by our camera system, centered at the table in the left figure. The omnidirectional view images were used for the evaluation after converting into grayscale.

two synchronized cameras with two fisheye lenses, capturing 2448×512 pixels images[2], as shown in Fig.4 (b), at 30.0 fps. In this paper, we ran our system offline to evaluate the recognition rates. The number of particles was set to 2,000 for each subject.

3.1 Labeling of facial expressions

Two other persons labeled all subjects with the dominant facial expression category at every frame in the video sequence. The labelers were denied the audio signal so labeling was based on just visual information. They were allowed to label a frame as 'smile' even if the participant was speaking. A reference expression at every frame was defined to be the union set: Correspondence between the two labelers yielded the final label. If their labels were divided into smile and laughter, the frame was labeled 'Mismatched'. Otherwise, if one of the two labels was 'Others', the other label was gave to the frame. Other all frames were labeled 'Others'. Examples of these labels and the resultant reference expressions are shown in Fig.5 and Fig.6. In addition, the labelers were also asked to assign labels indicating gaze direction.

3.2 Evaluation for facial expression recognition

The recognition rates were calculated as the ratio between the number of frames wherein the estimated expression matched the reference label to the total number of target frames. Moreover, the frames labeled 'Mismatched' were excluded in calculating the recognition rates. In this paper, the frames in the first third of the video sequence were used for training, and the other frames were for testing[3]. Table 1 shows the confusion matrix between facial expressions. Figure 5 shows the sequential recognition results together with two manual labels and their unions, or reference labels. In this figure, facial expressions in many frames are correctly recognized from the middle to end of the sequence, which are not included in the training data. Some random-noise-like misrecognition can be expected to be reduced by training that considers the transition of facial expression, $P(e_t|e_{t-1})$, from reference label data.

[2]Although the original image size of each camera is 2448×2048 pixels, only a horizontal strip 2448×512 pixels that covers the upper-body of meeting participants was stored.

[3]More properly, video sequence was divided into three groups, as each group has the same number of continuous sections, or the sections containing single facial expression label.

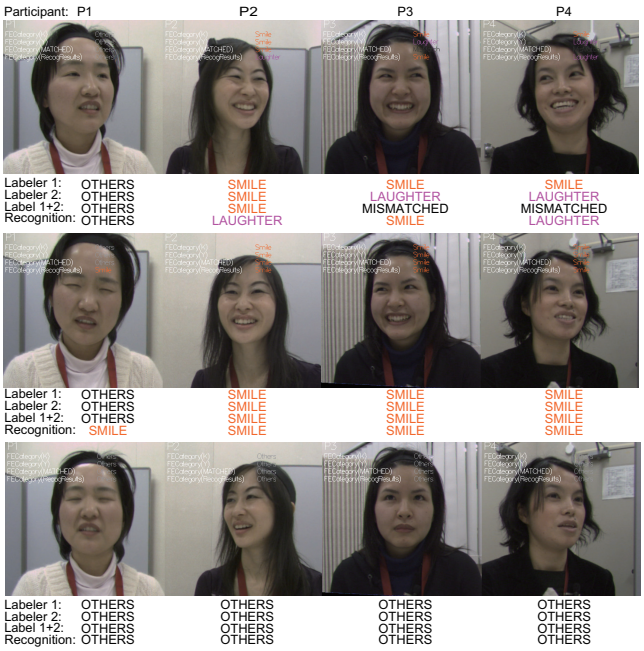


Figure 6: Sample resulting frames at frame 9426, 11754 and 12297 (top to bottom), respectively. The annotated labels given by two labelers (in the first two rows), the reference label, or the product of these two labels, (in the third row), and the recognition results (in the lowest row) are drawn in the image of each participant.

Table 1: Average confusion matrix of facial expressions: LBL and RCG denote reference label and recognition result, respectively.

LBL \ RCG	Smile	Laughter	Others
Smile	80.9	4.1	15.0
Laughter	8.0	83.4	8.6
Others	7.9	6.2	85.9

unit is (%)

Figure 6 shows the recognition results at three frames (frame numbers 9426, 11754, and 12297) in the sequence[4]. Some expressions are mistakenly recognized, i.e. smile of person P_2 at the top frame (9426) and 'Other' expression of person P_1 at the middle frame (11754). However, the border of these spontaneous expressions is quite ambiguous, as the labelers actually gave different labels for persons P_3 and P_4 at the top frame. Considering these ambiguities in manual labeling, the recognition rates are expected to be sufficiently high for globally identifying the frames where each participant was smiling or laughing.

3.3 Evaluation for interpersonal emotion estimation

Figure 7 shows the estimated interpersonal emotion network, together with that created from the manual labels

[4]A part of video sequence for the result is available from a supplemental movie.

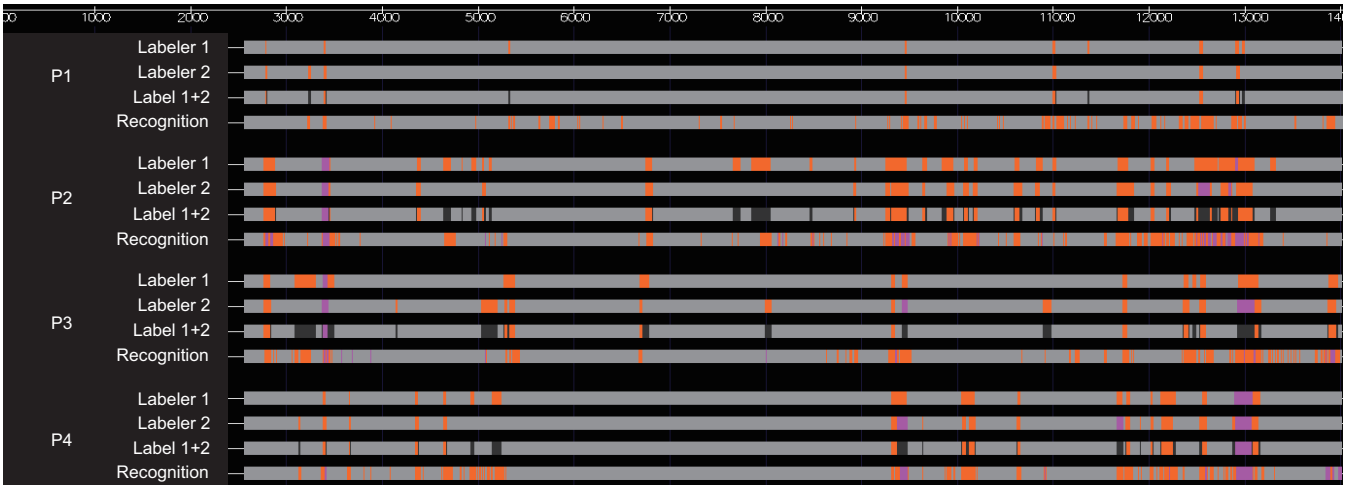
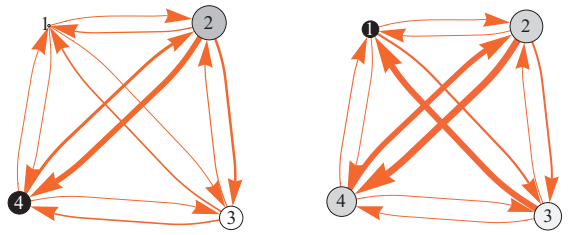


Figure 5: Sequential recognition results of facial expressions: Horizontal axis represents frame number, or time. Vertical axis shows the annotated labels given by two labelers (in the first two rows), the reference label, or the union of these two labels, (in the third row), and the recognition results (in the lowest row), for each of four participants. Colors denote expression categories: Orange: smile, pink: laughter, light gray: others, and dark gray: mismatched between labelers.



(a) Manual labels (b) Estimation

Figure 7: Interpersonal emotion networks created by manual labels (left) and estimation results (right). Number: participant. Thickness of arrow: amount of smiling, $S_{i \to j}$. Area of node: volume of received smiling, S_{IN}. Gray-level of node: volume of sent smiling, S_{OUT}. Moreover, the arrangement of the nodes represents the actual positioning of the participants.

of expression categories and their directions. The event of two people smiling at each other is indicated by the paired arrows. The people who often smiled are shown as large nodes with thick outgoing arrows. The people who were often smiled at by other participants are shown as dark nodes with thick incoming arrows.

The global trends of the network created by the proposed method resemble those yielded by the reference labels. The strongest link between persons P_2 and P_4 indicates that their relationship is closest among all pairs of the four participants. Although the largest node of person P_2 indicates that she smiled more often than others, her large variance among S_{OUT} demonstrates that she smiled selectively for person P_4. On the other hand, the smallest node, person P_1, indicates that she rarely smiled compared to the others.

However, person P_1 is mistakenly considered as the person who received the greatest number of smiles. The error comes mainly from error in estimating the direction of facial expression[5]. From the difference in the thickness of arrows between the two networks in Fig.7, we can see that the gazes of each person, especially person P_3, directed to the other participant sitting on the left or right side of the target person, and those averted are sometimes misrecognized as being directed to the participant in front of the target person. The reason seems to be that person P_3 frequently gazed at the other participant just by slightly rotating her head, as shown in Fig.6. Moreover, we confirmed that the interpersonal emotion network created with the recognized facial expressions and reference direction labels much more resembles that generated from manually-extracted expressions and their direction labels. Accordingly, improving the gaze estimation module is expected to enhance the quality of the interpersonal emotion network.

4. SUMMARY AND FUTURE WORK

We focused on smile as the indicator of interpersonal emotion in meetings, and recognized "how often who is smiling at whom". First, we proposed a new algorithm for the joint estimation of facial pose and expression in the framework of the particle filter. Its main feature is the automatic selection of interest points that can robustly capture small changes in expression against large head rotations. We then visualized interpersonal smiles as a graph structure, we call it the interpersonal emotional network; it indicates the emotional relationship among meeting participants. An evaluation using a four-person meeting captured by an omnidirectional video system suggested the effectiveness of the proposed method.

In the next step, we intend to extend the interpersonal emotion network by processing other facial expressions, e.g. wry smile, interested and thinking, as well as other modalities such as gesture, posture and utterance. For example, with regard to audio sources, the presence/absence of vo-

[5]The matching rates between the estimated directions and manual labels were 0.38, 0.79, 0.47, and 0.60 for P_1, P_2, P_3, and P_4, respectively.

calization is useful for more accurately discriminating smiles from laughter, or prosody for catching positive/negative feelings. We would also like to evaluate more estimated interpersonal emotion networks both qualitatively, e.g. by questionnaire, and quantitatively, e.g. co-occurrence of expressions, using a variety of data, i.e. different numbers of participants and conversational types such as cooperative/hostile discussions. Again, authors believe that automatically discovering the interpersonal emotions that evolves over time in meetings e.g. how each person feels about the others, or who is affectively influencing the others the most, is a promising and important research area.

5. REFERENCES

[1] M. Argyle. Bodily Communication – 2nd ed. Routledge, London and New York, 1988.

[2] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. J. Multimedia, 1(6):22–35, 2006.

[3] J. N. Cappella. Behavioral and judged coordination in adult informal social interactions: vocal and kinesic indicators. J. Personality and Social Psychology, 72(1):119–131, 1997.

[4] J. Cohn, L. Reed, T. Moriyama, J. Xiao, K. Schmidt, and Z. Ambadar. Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles. In Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, pages 129–138, 2004.

[5] F. Dornaika and F. Davoine. Simultaneous facial action tracking and expression recognition in the presence of head motion. Int'l J. Computer Vision, 76(3):257–281, 2008.

[6] B. Fasel and J. Luettin. Automatic facial expression analysis: Survey. Pattern Recognition, 36:259–275, 2003.

[7] D. Gatica-Perez. Analyzing group interactions in conversations: a review. In Proc. IEEE Int'l Conf. Multisensor Fusion and Integration for Intelligent Systems, pages 41–46, 2006.

[8] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific Active Appearance Models. Image and Vision Computing, 23(11):1080–1093, 2005.

[9] A. Ito, X. Wang, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In Proc. Int'l Conf. Cyberworlds, pages 437–444, 2005.

[10] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations from nonverbal activity cues. IEEE Trans. Audio, Speech, and Language Processing, 17(3):501–513, 2009.

[11] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. Int'l J. Computer Vision, 83:178–194, 2009.

[12] D. Mikami, K. Otsuka, and J. Yamato. Memory-based particle filter for face pose tracking robust under complex dynamics. In Proc. IEEE Conf. Computer Vision and Pattern Recognition, pages 999–1006, 2009.

[13] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In Proc. ICMI, pages 257–264, 2008.

[14] K. Otsuka, H. Sawada, and J. Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations. In Proc. ICMI, pages 255–262, 2007.

[15] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In Proc. CHI, pages 1175–1180, 2006.

[16] M. Pantic and M. Bartlett. Machine Analysis of Facial Expressions. I-Tech Education and Publishing, 2007.

[17] S. Petridis and M. Pantic. Audiovisual laughter detection based on temporal features. In Proc. ICMI, pages 37–44, 2008.

[18] J. Russell, J. Bachorowski, and J. Fernandez-Dols. Facial and vocal expressions of emotion. Annual Review of Psychology, 54:329–349, 2002.

[19] B. Schuller, R. Muller, B. Hornler, A. Hothker, H. Konosu, and G. Rigoll. Audiovisual recognition of spontaneous interest within conversations. In Proc. ICMI, pages 30–37, 2007.

[20] R. Stiefelhagen. Tracking focus of attention in meetings. In Proc. ICMI, pages 273–280, 2002.

[21] Y. L. Tian, T. Kanade, and J. Cohn. Facial expression analysis. Springer, 2005.

[22] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In Proc. ICMI, pages 38–45, 2007.

[23] J. A. R. A. M. van Hooff and S. Preuschoft. Animal social complexity: Intelligence, culture, and individualized societies, chapter Laughter and smiling:The intertwining of nature and culture, pages 260–287. Harvard University Press, 2003.

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proc. IEEE Conf. Computer Vision and Pattern Recognition, pages 511–518, 2001.

[25] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell., 31(1):39–58, 2009.