Analyzing Empathetic Interactions based on the Probabilistic Modeling of the Co-occurrence Patterns of Facial Expressions in Group Meetings

Shiro Kumano*, Kazuhiro Otsuka, Dan Mikami, and Junji Yamato
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya, Atsugi, Kanagawa, Japan
*kumano.shiro@lab.ntt.co.jp, kumano@eye.brl.ntt.co.jp

***Abstract*— This paper presents a novel research framework for the estimation of emotional interactions produced between meeting participants. The types of emotional interaction targeted in this paper are empathy, antipathy, and unconcern. We define here emotional interaction as a brief contiguous event wherein a pair exchange emotional messages via verbal and non-verbal behaviors. As the key behaviors, we focus on facial expression and gaze, because their combination realizes the rapid and directed transmission of a large number of emotional messages. We assume that there is a strong link between the emotional interaction and the participants' facial expressions that occur simultaneously with the type of the emotional interactions. Based on this assumption, we build a probabilistic model that represents a hierarchical structure involving the emotional interactions, facial expressions and other behaviors including utterance and gaze direction. Using this model, the type of emotional interaction is estimated from interpersonal gaze directions, facial expressions, and utterances. Our estimation is based on the Bayesian approach, and uses the Markov chain Monte Carlo method to approximate joint posterior probability distributions of the emotional interaction and model parameters present within the observed data. An experiment on four-party conversations demonstrates the promising effectiveness of the proposed method.**

I. INTRODUCTION

Face-to-face conversation is the primary way of sharing information, understanding others' emotion, and making decisions in social life. Accordingly, multimodal meeting analysis has been acknowledged as an emerging research area and intensive efforts have been made to analyze meetings as introduced in [1]. Meetings could be more interesting, meaningful and productive, if a system were available that could carefully support the meetings, like an experienced facilitator, by reading a wide range of meeting conditions, such as turn taking/floor control shift [2], dominance person [3], person's emotion, e.g. six basic emotions and interpersonal emotions, and the relationship between the participants. However, among these meeting conditions, this emotional aspect has never been comprehensively addressed in the field of automatic meeting analysis.

To deeply understand meetings, it is critical to grasp how consensus as well as the relationship among participants is established through the exchanges of emotional messages. The basic units of the emotional message exchange in meetings are empathy, unconcern, and antipathy; empathy often indicates an agreement and a close relationship. Empathy causes behavioral coordination[1][4], and has been termed motor mimicry [5] or the Chameleon effect [4]. Empathy often produces the coordination of positive emotional behaviors, e.g. smile to smile [4]. The coordination of negative emotional behaviors may occur, e.g. painful face when someone suffering [5]. On the other hand, behavioral incoordination is thought to be a sign of antipathy or unconcern.

Among the participant behaviors related to empathy, unconcern, or antipathy, the combinations of facial expression (FE) and gaze are especially important, because they realize the rapid and directed transmission of a large number of emotional messages. Facial expression plays a major role in conveying emotional messages [6], [7]: e.g. smiling indicates positive feelings, cooperative partners [8] and cooperative alliances [9], and builds/maintains intimacy or rapport [10], while negative FEs such as frowns convey sadness, puzzlement, etc. Neutral expressions, including the absence of any reactive expressions, may be seen as an unconcern or antipathetic response to a positive FE.

We note that gaze has several functions [11], [12], including monitoring and triggering others' reactions. These functions are especially important to exchange emotional messages. Monitoring is vital to reading the other's FEs and inferring his/her emotion from them. Gaze can trigger a gaze shift of the gazee toward the gazer: If person X is looking at person Y, then Y notices X's gaze and turns his/her gaze upon X. At this moment, they establish in mutual gaze or eye contact, where emotional messages conveyed by FEs are rapidly exchanged or shared.

We use the term *emotional interaction* to refer to brief events, where a pair of people exchange emotional messages through a combination of FE and gaze behavior. An emotional message may either be shared, ignored, or refused. In short, the emotional interaction explains, "*who is sharing empathy/antipathy with whom*". We provide a brief explanation with typical examples. Suppose person X and Y are feeling empathy for each other, but it has not been shared between them yet. Now, X and Y are looking at each

[1]This paper supports the assumption that empathy is *the cause of* behavioral coordination, though there are also opposite arguments that empathy is *the result of* behavioral coordination [4].

other. First, X smiles at Y while hoping that Y will return the feeling, and Y receives X's smile. Y returns the smile to X in order to show that Y is feeling empathy for X. Last, X realizes Y's feeling; At this moment, the empathy is shared between them. Y may return negative FEs to X, when X is demonstrating emotional pain with negative FEs. In contrast, if the emotional feeling between them is antipathetic, Y may return an FE opposite to X's FE: e.g. disgust expression for smile, derisive smile for sad expressions. If X and Y are unconcerned with each other, they will often show neutral expressions regardless of the partner's expression, or even make little attempt to look at each other.

Based on the aforementioned assumption that the co-occurrence of FEs for each pair of participants depends on the type of emotional interaction and gaze state between them, this paper first proposes a novel research framework of estimation of emotional interactions presented between meeting participants. We build a probabilistic conversation model that can represent a hierarchical structure, where emotional interaction evolves through Markovian transitions, and gaze and facial expression governed by the emotional interaction, and gaze, facial expression and utterance are observed in an video frame or image. Each pair at each time takes one of three types of emotional interaction, empathy, unconcern and antipathy. We model the co-occurrence pattern of FEs with a frequency matrix, we call *facial expression co-occurrence matrix* and denote herein as *FE co-occurrence matrix*, where each component describes how likely the corresponding combination of FEs in the participant pair is to occur. Each type of emotional interaction and gaze state has a different FE co-occurrence matrix. Each participant is assumed to probabilistically exhibit an FE according to the co-occurrence pattern at the time.

Using this model, the emotional interaction are estimated from interpersonal gaze directions, facial expressions, and utterances. We employ Bayesian estimation to approximately calculate the joint posterior probability distribution of emotional interaction, facial expression and model parameters given by the observations, by utilizing a Markov chain Monte Carlo (MCMC) method called the Gibbs sampler [13], which has an advantage when dealing with complex models.

The remainder of this paper is organized as follows. First, in Section II, related works are introduced. Next, Section III describes our FE co-occurrence matrix. Next, in Section IV and Section V our proposed conversation model and the estimation achieved by using a Gibbs Sampler are explained, respectively. Section VI describes the results of an experiment. Finally, a summary and future work are given in Section VII.

II. Related works

To the best of our knowledge, no attempt has been made to automatically estimated emotional interaction in group meetings, in spite of the importance of emotional interaction, e.g. see the recent comprehensive reviews of the automatic detection/recognition of human social behaviors [14], [15]. This section briefly overviews the existing scientific or engineering works related to our research from the following four viewpoints: facial expression recognition, estimation of interpersonal estimation, estimation of the state of meeting group at high level, and estimation of social network or Sociometry.

First, the targets of facial expression recognition are shifting from the previous deliberate and exaggerated basic FEs, e.g. found in [16], [17], [14], to spontaneous ones [18], [19], [20]. However, the existing methods still focus on just the person who is showing facial expressions. Furthermore, few studies target subtle spontaneous FEs, such as wry smile, exhibited in the social multiparty meeting setting, and little attention has been paid to the communicative function of FEs, i.e. transmitting emotional messages.

Second, in [21], the authors proposed a method to automatically visualize a network of interpersonal emotion between participants. The method first detects smiles individually, and then calculates the number of smiles exchanged between participants by combining gaze behaviors, i.e. the method recognizes "who is smiling at whom, when, and how often". This method is based on the assumption that an FE directed to another person straightforwardly indicates the emotion of the sender to the receiver. However, negative FEs of a pair, for example, may co-occur, if they are negatively emphasizing a conversation topic, e.g. smoking. Such emotions toward a third-party or object is out of scope of the method.

Third, few studies have attempted to understand meeting states at relatively higher-levels, e.g. conversation regimes such as monologue or dialogue [22], and the dominant person in meeting [3]. Of particular note, [22] proposed a hierarchically structured conversation model, linking regime, interaction and behavior layers. However, the emotional aspect of meetings was basically not addressed.

Last, there have been some recent attempts on inferring a social network, also called Sociometry, where persons are represented as nodes and each link between them indicates the presence of a relationship between them. The social networks were inferred by directly linking them with low-level information, such as physical proximity [23] and low-level image features [24]. However, little attention has been paid to the aspects of short-term emotional state such as empathy in this research area.

III. Co-occurrence pattern of Facial Expressions in Emotional Interactions

We performed preliminarily quantitative analysis on the relationship between emotional interaction, facial expression (FE) and gaze by using actual data taken from four four-party conversations as explained in VI. We here briefly verify our assumption that the co-occurrence pattern of FEs, describing how the FEs of paired participants co-occur, depends on the type of emotional interaction and gaze state between them.

A. Definition of emotional interaction, gaze and facial expression

The emotional interaction in this paper is categorical defined for each pair of persons. Three categories of emotional

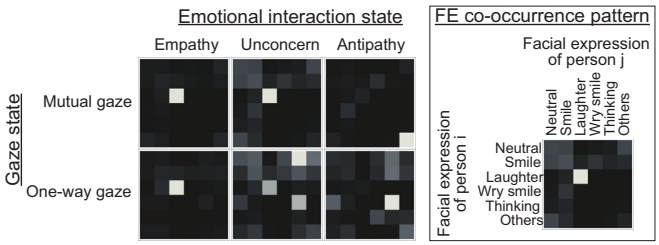

Fig. 1. FE co-occurrence matrix for each emotional interaction and gaze state: Each 6×6 matrix describes the FE co-occurrence matrix between person i and person j. Brighter component means higher frequency. In one-way gaze, person i is a gazer, i.e. only person i is looking at person j.

interaction are set in this paper: *empathy*, *unconcern*, and *antipathy*. The emotional interaction is an undirected event between a pair of participants, wherein a pair exchange emotional messages via verbal and non-verbal behaviors.

An facial expression is a categorical state defined for each individual. FEs are grouped into three categories in this paper: *neutral*, *smile*, *laughter*, *wry smile*, *thinking*, and *others* expressions. "Others" category contains expression of embarrassing, disgusting, surprising etc.

The gaze patterns takes one of three states: *mutual gaze*, *one-way gaze* and *mutually averted gaze*. Mutual gaze, or eye contact, means the state wherein both persons in the pair are looking at each other. One-way gaze means the state wherein one of the pair is looking at the partner, though the partner is not. Mutually averted gaze means the state wherein neither is looking at the other; in this paper, by focusing on visual behaviors, we assume there are no significant interactions between a pair in their mutually averted gaze.

B. Facial expression co-occurrence matrix

The co-occurrence pattern of FEs is modeled by a frequency matrix, which we call the *FE co-occurrence matrix*. Fig.1 shows the facial expression (FE) co-occurrence matrices for each emotional interaction and gaze pattern, calculated from data extracted from our recorded conversations with hand labels, described in VI. Moreover, the FE co-occurrence matrices for mutual gaze are symmetric, while those for one-way gaze are asymmetric because of the physical asymmetry of the one-way gaze.

The well separated patterns between FE co-occurrence matrices in Fig.1 supports the validity of our assumption that the FE co-occurrence pattern depends on the type of emotional interaction and gaze state. Furthermore, this tendency basically agrees with existing psychological assessments: a) In the empathetic interactions, the co-occurrence of positive expressions, i.e. smile and laughter, is quite frequent [4]. b) Mutual gaze enhances behavioral coordination [5]: e.g. the FE co-occurrence matrices of mutual gaze are biased more to the co-occurrence of positive FEs than those of one-way gaze. Other intuitively reasonable characteristics can also be found: c) In the antipathetic interactions, "wry smile", "thinking", and "others" expressions are shown by one or both of the pair of participants. d) In the unconcern interactions, the co-occurrence of neutral expression and other expression are significant.

C. Factorization of FE co-occurrence matrix

The FE co-occurrence matrices in this paper are factorized by dividing their original co-occurrence matrices by the marginal distribution, or composition rate, of the facial expressions of each participant. That is, the FE co-occurrence matrices in Fig.1 were obtained by first counting the frequencies of each case of FE co-occurrence for each person pair in the data, then dividing each of them by total frequency of each FE of each participant, i.e. how frequently each FE category was exposed by a target person in the data, and last averaging them. The factorization can well reveal the differences in the patterns of FE co-occurrence matrices between emotional interactions, because it reduces the impact of the frequency of each FE category, which depends on the conversation conditions, such as participants' personality, as well as meeting type and conversation theme. As the psychological assessment that people often smile in the presence of an observer [9], the FE co-occurrence of smiles actually occurred often in our data regardless of the type of emotional interaction.

IV. Conversation model

This section describes our probabilistic model that represents a hierarchical structure involving the emotional interactions and behaviors, i.e. gaze patterns, facial expressions, and utterances. Using this model, emotional interaction is estimated from interpersonal gaze directions, facial expressions, and utterances.

A. Model structure

To model the relationship between emotional interaction and the participants' behaviors for N-party meeting, this paper employs a hierarchical dynamic Bayesian network (DBN); a discrete random process at a higher level evolves through Markovian transitions, and the lower levels are governed by higher levels. Here, the high-level process corresponds to the type of emotional interaction and the lower one corresponds to participants' behaviors including FE and gaze patterns.

Fig.2 shows a graphical representation of our hierarchical DBN for estimating emotional interaction and FE from audio-visual signals for discrete temporal interval $t = 1, \cdots, T$: nodes represent variables and edges represent dependencies between variables. Our model consists of four discrete random variables: emotional interaction $\{E_t\}_{t=1}^T$, facial expression $\{F_t\}_{t=1}^T$, gaze pattern $\{X_t\}_{t=1}^T$, and utterance $\{U_t\}_{t=1}^T$. Emotional interaction and gaze state between a pair are assumed to affect their facial expressions. Note that their relationship is modeled with the FE co-occurrence matrices, described in III. The type of emotional interaction also influences gaze state. This describes the tendency found in our preliminary experiment that sympathetic and antipathetic interactions are likely to cause mutual gazes, while unconcern interactions often produce one-way gazes. Furthermore,

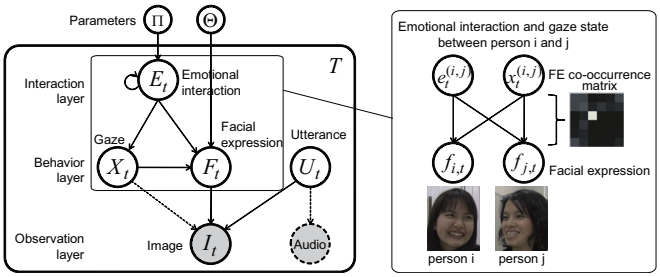

Fig. 2. Graphical representation of the proposed model: The nodes in gray, or auditory and visual signals I, are observations. Moreover, the states of gaze X and utterance U (and facial expression F) are, in this paper, assumed to be given. In this case, the dotted node and arrows can be omitted.

image of each person is assumed to be generated according to his/her own FE and utterance state at the time; The utterance state affects mainly the shape of the mouth.

The set of emotional interaction of all participant pairs at time t is denoted as $E_t = \{e_t^{(i,j)}\}_{(i,j)\in \boldsymbol{r}}$, where $e^{(i,j)} \in \boldsymbol{e} = \{1, \cdots, N_e\}$ denotes the emotional interaction between the pair of i and j, (i,j), and $\boldsymbol{r}$ represents the set of $N \times (N-1)/2$ combinations of participant pairs. N_e is the number of emotional interaction. In this paper, $e \in \{$*"empathy"*, *"unconcern"*, *"antipathy"*$\}$, $N_e = 3$. The set of FE states is denoted as $F_t = \{f_{i,t}\}_{i=1}^N$, where $f_i \in \boldsymbol{f} = \{1, \cdots, N_f\}$ is FE state of person i. N_f is the number of FE categories In this paper, $f \in \{$*"neutral"*, *"smile"*, *"laughter"*, *"wry smile"*, *"thinking"*, *"others"* $\}$, $N_f = 6$. The set of gaze patterns of all pairs at time t is denoted as $X_t = \{x_t^{(i,j)}\}$, where the gaze patterns between pair (i,j) at time t, $x_t^{(i,j)}$, takes one of three states: $\{$*"mutual gaze"*, *"one-way gaze"*, *"mutually averted gaze"*$\}$. The utterance pattern is denoted as $U_t = \{u_{i,t}\}_{i=1}^T$, where the utterance state $u_{i,t}$ indicates whether person i is making utterance, or not at time t.

Observable variables $Z_{1:T}$ basically consist of the sequences of images $I_{1:T}$ and audio signals. If already obtained by using other method(s), all or some of the behaviors can be added to the observations. Section VI demonstrates the experiments gained under two conditions: (Cond. A) the estimation of emotional interaction from $Z_{1:T} = \{X_{1:T}, F_{1:T}\}$, and (Cond. B) the joint estimation of emotional interaction and FE from $Z_{1:T} = \{X_{1:T}, U_{1:T}, I_{1:T}\}$. IV-B and V describe Cond. B. In Cond. A, the terms that contain only the symbols of observations in the each equation can be folded into the normalization constant.

B. Joint probability distribution

Based on the proposed model, the problem of this paper is to estimate the emotional interaction sequence $E_{1:T}$, FE sequence $F_{1:T}$, and model parameters φ from observations $Z_{1:T}$. We employ a Bayesian approach to estimate the joint posterior distribution of all unknown variables for given measurements, $p(E_{1:T}, F_{1:T}, \varphi | Z_{1:T})$. In Bayesian analysis, a priori knowledge about the model is represented as the prior distributions of model parameters, $p(\varphi)$.

The joint probability distribution is defined as

$$\begin{aligned} &p(E_{1:T}, F_{1:T}, X_{1:T}, U_{1:T}, I_{1:T}, \varphi) \\ &:= p(\varphi) P(E_{1:T}|\varphi) P(X_{1:T}|E_{1:T}, \varphi) \\ &\quad P(F_{1:T}|E_{1:T}, X_{1:T}, \varphi) P(I_{1:T}|F_{1:T}, U_{1:T}, \varphi). \end{aligned} \quad (1)$$

Hereafter, the set of parameters φ is omitted for notational simplicity unless necessary.

Assuming the first-order Markov process and independency for all pairs, the prior probability of emotional interaction is decomposed as

$$P(E_{1:T}) := \prod_{t=1}^{T} \prod_{(i,j)\in \boldsymbol{r}} P(e_0^{(i,j)}) P(e_t^{(i,j)} | e_{t-1}^{(i,j)}), \quad (2)$$

where $P(e_0^{(i,j)}) = \pi_{0,e'}^{(i,j)}$ and $P(e_t^{(i,j)} = e' | P(e_{t-1}^{(i,j)} = e) = \pi_{e,e'}^{(i,j)}$ are an initial probability and a transition probability of emotional interaction. Both of the probabilities are constant over time, where $\sum_{e'\in \boldsymbol{e}} \pi_{\cdot,e'} = 1$. The likelihood of emotional interaction for gaze state, $P(X_{1:T}|E_{1:T})$, is defined as the product of that of each pair and time, i.e.

$$P(X_{1:T}|E_{1:T}) := \prod_{t=1}^{T} \prod_{(i,j)\in \boldsymbol{r}} P(x_t^{(i,j)} | e_t^{(i,j)}). \quad (3)$$

Each pair has its own parameter, so the likelihood parameters are denoted as $P(x_t^{(i,j)} = x | e_t^{(i,j)} = e') = \pi_{x,e'}^{(i,j)}$. These parameters related to emotional interaction, $\boldsymbol{\pi}_{\cdot} = \{\pi_{\cdot,e'} | e' \in \boldsymbol{e}\}$, are assumed to follow independent Dirichlet distributions.

The conditional probability of FE of all participants given by emotional interaction and gaze states is defined to be the product of the FE prior probabilities for all participants and the FE co-occurrence matrix described in III:

$$\begin{aligned} P(F_{1:T}|E_{1:T}, X_{1:T}) &:= \prod_{t=1}^{T} \prod_{i=1}^{N} P(f_{i,t}) \cdot \\ &\prod_{t=1}^{T} \prod_{(i,j)\in \boldsymbol{r}} \mathcal{M}(f_{i,t}, f_{j,t} | e_t^{(i,j)}, x_t^{(i,j)}), \end{aligned} \quad (4)$$

where $P(f_{i,t})$ denotes the prior distribution of FE of person i at time t. The FE co-occurrence matrix $\mathcal{M}$ is prepared for each emotional interaction and gaze state; the number of the FE co-occurrence matrices is $N_e \times N_g$. The size of each FE co-occurrence matrix is $N_f \times N_f$. The FE co-occurrence matrices for mutually averted gaze are fixed to be uniform matrices in this paper. Each pair and person has these model parameters. The parameters of the prior distribution are $P(f_i = f') = \theta_{i,f}$, where $\boldsymbol{\theta}_{\cdot} = \{\theta_{\cdot,f'}\}_{f'\in \boldsymbol{f}}$ are assumed to follow independent Dirichlet distributions. The parameter of each element (f, f') of each $N_f \times N_f$ FE co-occurrence matrix, $\mathcal{M}_{f,f'} (> 0)$, is $\mathcal{M}(f_i = f, f_j = f' | e^{(i,j)} = e, x^{(i,j)} = x) = \gamma_{(i,j),e,x,ff}$, where $\boldsymbol{\gamma}_{\cdot} = \{\gamma_{\cdot,ff}\}_{ff\in\{1,\cdots,N_f\times N_f\}}$ are also assumed to follow independent Dirichlet distributions.

The likelihood of facial expression and utterance for observed image or video frame, $P(I_{1:T}|F_{1:T}, U_{1:T})$, is assumed

to be independent between participants:

$$P(I_{1:T}|F_{1:T}, U_{1:T}) = \prod_{t=1}^{T}\prod_{i=1}^{N} P(I_t|f_{i,t}, u_{i,t}). \quad (5)$$

This paper defines the likelihood as $P(I_t|f_{i,t}, u_{i,t}) := P(\mathrm{FER}_{i,u_{i,t}}(I_t)|f_{i,t})$, where $\mathrm{FER}_{i,u}(I)$ is a facial expression recognizer for person i in utterance state u; it returns an estimate of facial expression category, $\tilde{f}$, from incoming image I. Details of the FE recognizer are described in VI-B. The parameters of the distribution are $P(\mathrm{FER}(\cdot) = f|f_i = f') = \theta_{i,f,f'}$, where $\boldsymbol{\theta}_\cdot = \{\theta_{\cdot,f'}\}_{f'\in\boldsymbol{f}}$ are assumed to follow independent Dirichlet distributions.

In summary, the full model parameters mentioned above are written as $\varphi = \{\Pi, \Theta\}$. The model parameters related to emotional interaction, Π, are denoted as $\Pi = \{\boldsymbol{\pi}_0^{(i,j)}\}_{(i,j)} \cup \{\boldsymbol{\pi}_e^{(i,j)}\}_{(i,j),e} \cup \{\boldsymbol{\pi}_x^{(i,j)}\}_{(i,j),x}$. FE-related model parameters, Θ, are denoted as $\Theta = \{\boldsymbol{\theta}_i\}_i \cup \{\boldsymbol{\gamma}_{(i,j),e,x}\}_{(i,j),e,x} \cup \{\boldsymbol{\theta}_{i,f}\}_{i,f}$. The prior $p(\varphi)$ is defined as the product of that of each of the parameters.

V. Bayesian estimation via Gibbs sampling

Based on the conversation model proposed in IV, the problem here is to estimate the joint posterior probability distribution of the emotional interaction sequence $E_{1:T}$, FE sequence $F_{1:T}$, and model parameters φ given by observations $Z_{1:T}$, $p(E_{1:T}, F_{1:T}, \varphi|Z_{1:T})$. This study utilizes the Gibbs sampler [13], a variant of the Markov chain Monte Carlo (MCMC) method, due to its advantages in dealing with complex models. The Gibbs sampler repeatedly generates random samples from the full conditional posterior distributions of each unknown variable, which constitute a Markov chain whose invariant distribution equals the desired joint posterior. The joint posterior distribution is approximated by a set of random samples after the Markov chain has converged, and is used to calculate statistics. This study employs natural conjugate prior distributions [25] for mathematical convenience. As the conjugate prior of all parameters, we use independent Dirichlet distributions, which are commonly used as the prior of discrete random variables.

A. Full conditional posterior distributions

The full conditional posterior distribution of each variable has the same function form as its priors, since natural conjugate priors are used. It is proportional to a distribution of only those components that are unrelated to the target variable from the joint distribution in (1); other components can be folded into the normalization constant.

Emotional interaction type at time step t, $e_t^{(i,j)}$, is sampled according to its full conditional probability, as given by

$$P(e_t^{(i,j)}|E_{1:T}\backslash e_t^{(i,j)}, F_{1:T}, \varphi, Z_{1:T}) \propto P(e_t^{(i,j)}|e_{t-1}^{(i,j)}) P(e_{t+1}^{(i,j)}|e_t^{(i,j)}) P(x_t^{(i,j)}|e_t^{(i,j)}) \mathcal{M}(f_i, f_j|e_t^{(i,j)}, x_t^{(i,j)}) \quad (6)$$

Facial expression, $f_{i,t}$, can be sampled in a similar manner.

Each of the parameters, which follows a Dirichlet distribution, can be updated by adding the total number of time steps at which the target event occurred to current time.

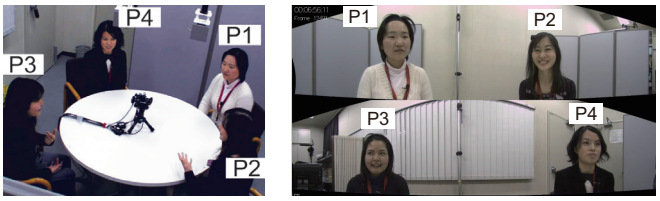

Fig. 3. Example scene of the conversation. Left: Overall view taken by an extra handy camera. Right: Omnidirectional view captured by our camera system, centered at the table in the left figure. The omnidirectional view images were used for the evaluation after converting into grayscale.

For example, the parameter of the transition probability of emotional interaction, $\pi_{e,e'}^{(i,j)}$, is updated as $\pi_{e,e'}^{(i,j)} \leftarrow \pi_{e,e'}^{(i,j)*} + n_{e,e'}^{(i,j)}$, where $\pi_{e,e'}^{(i,j)*}$ is a hyper parameter of the parameter, and $n_{e,e'}^{(i,j)}$ denotes the total number of time steps where emotional interaction between pair (i,j) changes from e to e' in between adjacent frames.

B. Estimates

After the iterations terminate, statistics are calculated from the samples $\{E_{1:T}^{(q)}, F_{1:T}^{(q)}, \varphi^{(q)}\}_q$ for iteration steps $q = M'$ to M to ensure convergence. The posterior probability distributions of emotional interaction are calculated as: $P(e_t^{(i,j)}|Z_{1:T}) \approx \sum_{q=M'}^{M} \delta_e(e_t^{(i,j)(q)})/(M - M' + 1)$, where δ is an indicator function as in $\delta_\xi(\xi') = 1$ if $\xi = \xi'$, otherwise $\delta_\xi(\xi') = 0$. The maximum a posterior (MAP) estimates of emotional interaction are obtained as: $\hat{e}_t^{(i,j)} = \arg\max_{e_t^{(i,j)}\in\boldsymbol{e}} P(e_t^{(i,j)}|Z_{1:T})$. The MAP estimates of FE, $\hat{f}_{i,t}$, are obtained in a similar manner. For the parameters, the minimum mean-squared error estimates are calculated: e.g. for parameter $\pi_\cdot$, $\pi_\cdot = \sum_{q=M'}^{M} \pi_\cdot^{(q)}/(M - M' + 1)$.

VI. Experiments

This section describes qualitative and quantitative evaluations of how closely the emotional interactions estimated from observations of recorded data of group meetings by the proposed framework are to their hand labels.

A. Data

This paper targets four-person group conversations. The participants were four women in the same age bracket, seated as shown in Fig.3. They had not met before the experiment. The participants were instructed to hold discussions about the following four topics: "Who is more beneficial, men or women?" (called C1), "Should people marry or not?" (C2), "Should smoking be fully prohibited in public spaces?" (C3), and "Is marriage and romantic love the same or different?" (C4). They were asked to try to reach a conclusion as a group on each discussion topic within eight minutes. The discussions were held on the same day. Each conversation was captured by a tabletop sensing device for roundtable meetings [26]; it consists of two synchronized cameras with two fisheye lenses, capturing $2448\times(512\times2)$ pixel images at 30 fps, and a triangular microphone array. The frame lengths of targeting data for C1-C4 were 14580, 12480, 11450 and 15760, respectively, and ranged from 6.4 to 8.8 min.

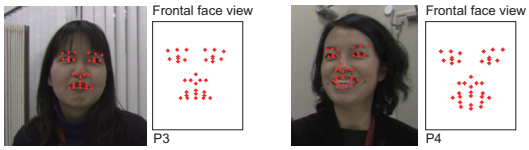

Fig. 4. Example tracking results of facial landmark points.

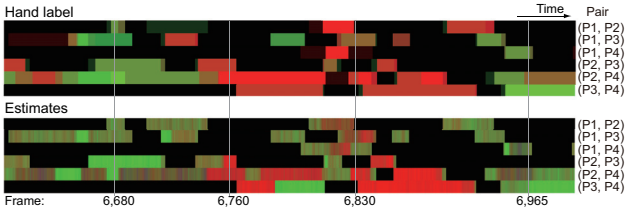

Fig. 5. Example sequence of emotional interaction of frame no. 6600-7000 ($\approx$ 13 sec) from the frames in (upper) those labeled by humans, and (lower) those estimated by the proposed method (Cond. A). Horizontal axis represents frame number, or time. Colors denote the frequency or probabilities of emotional interactions. R, G, and B components correspond to the frequency/probability of empathy, unconcern, and antipathy, respectively. Pure reds indicate that all labelers gave empathy label or the estimated type of interaction is empathy with high certainty. Mixed colors indicate ambiguous labeling or dispersed posterior distributions.

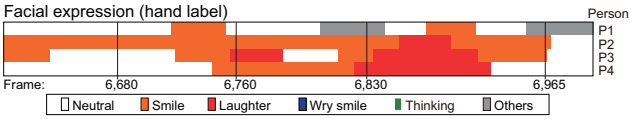

Fig. 6. Hand labels of facial expression: Horizontal axis equals to that of Fig.5. Colors denote categories of facial expression.

B. Settings for Gibbs sampling

The proposed method is evaluated in two conditions: the estimation of emotional interaction with $Z_{1:T} = \{X_{1:T}, F_{1:T}\}$ (Cond. A) and the joint estimation of emotional interaction and FE with $Z_{1:T} = \{X_{1:T}, U_{1:T}, I_{1:T}\}$ (Cond. B). The detection of gaze direction and utterance behavior of each participant lies outside the focus in this paper. Thus, these states were manually derived. To automate these processes, any feasible method can be utilized: for example, [27] for gaze tracker, and [28] for voice activity detector.

For a facial expression recognizer, any facial expression recognizer that inputs image I, can be utilized in our framework. This part also lies outside our focus. In this paper, we utilize Support Vector Machines (SVMs). These SVMs were individually trained for each participant based on the leave-one-conversation-out cross-validation scheme. Two types of SVMs were created: one is for utterance, and the other is for silence. Their inputs are eight geometrical features of facial landmark points, as shown in Fig.4, which are defined in [29], [30], i.e. distances of two points and angles between three points in eyebrow and mouth regions. To obtain the configuration of the facial landmark points in each image, we utilize a commercial face tracker[2].

Hyperparameters for prior distributions were empirically set to be the values trained by using all conversation data. The hyperparameters were constant for all pairs and each person, and all conversations; the corresponding parameters specific to the pair or person are estimated by Gibbs sampling. Estimation results were obtained after $M = 800$ iterations of Gibbs sampling ($M' = 600$). The number of iterations was chosen experientially by confirming the convergence of the estimates.

C. Labeling of emotional interaction and behaviors

Five labelers labeled all pairs with the emotional interaction at every frame in the video sequence without accessing to the audio signals. One of the labelers also labeled all subjects with FE category and gaze direction. Another labeler was asked to assign labels indicating utterance or silence. None of the labelers were conversation participants. As to emotional interaction, the labelers were asked to select one of the five labels, "strong empathy" $(+2)$, "weak empathy" $(+1)$, "unconcern" (0), "weak antipathy" (-1), and "strong antipathy" (-2). If there were no significant interactions between them, the labelers were allowed to use "no interaction". Examples of these labels are shown in Fig.5 and Fig.6. Note that all these labels were used for training of hyperparameters and evaluating estimations, but as for the estimation, only the labels of gaze direction and utterance were used as observations. Moreover, all frames exhibiting mutually averted gaze, decided by using the gaze labels, were automatically considered to be "no interaction" in advance. The frames labeled with "no interaction" were removed when evaluating estimation performance.

The total frequency of each emotional interaction label in mutual gaze or one-way gaze was 15.2[%] (strong empathy), 29.2[%] (weak empathy), 52.8[%] (unconcern), 2.5[%] (weak antipathy), and 0.3[%] (strong antipathy). The frequency of each FE was 36.5[%] (neutral), 52.4[%] (smile), 3.0[%] (laughter), 0.3[%] (wry smile), 3.7[%] (thinking), and 4.1[%] (others). Given the conversation topics, empathetic interactions and smiles were frequent, while the antipathetic interactions and others FEs were rare.

D. Qualitative evaluation of the estimates of emotional interaction

Fig.5 shows an example of sequential estimation results of emotional interaction estimated by using the proposed method (Cond. A), together with those of the hand labeles. The similarity of the color patterns between them suggests that the proposed method can estimate the emotional interaction not only with its type but also with an ambiguity similar to that of human labelers; The frames drawn in mostly red or green denote that most or all of the labelers assigned empathy or unconcern to these frames, or the method assigned high posterior probability to the type. Mixed colors mean that the labelers' hand labels disagreed, or that the method created wide probability distributions indicating that the estimation

[2]*FaceAPI*, Seeing Machines: http://www.seeingmachines.com/product/faceapi/.

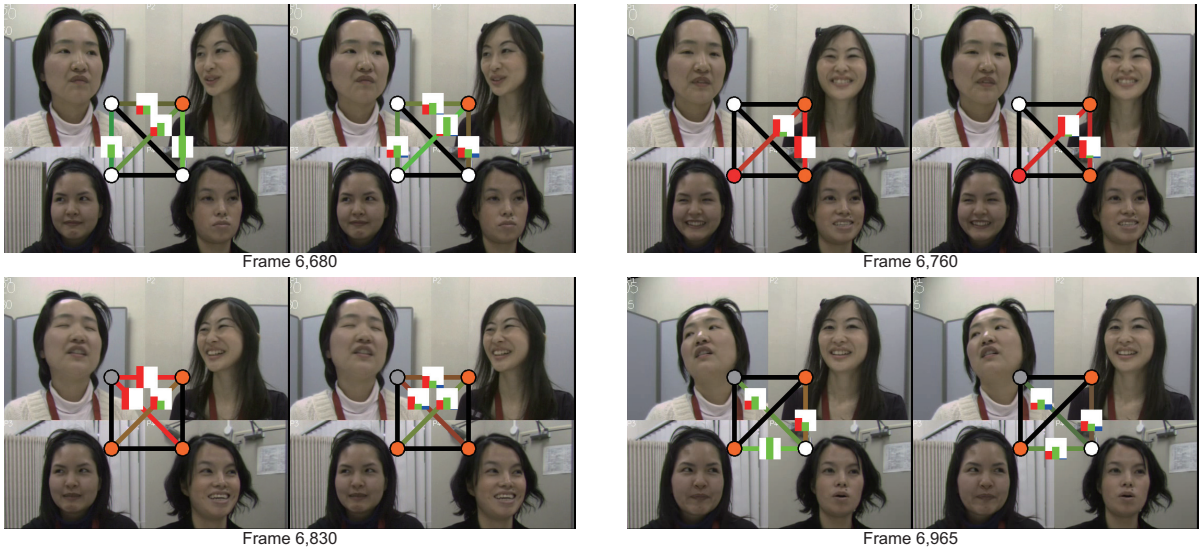

Fig. 7. Four snapshots of the scene, hand labels and estimation results (Cond. A) in the frames in Fig.5: The human labels and estimation results are illustrated as a graph structure that overlays on the participants. Left and right figures denote human labels and estimations, respectively. Each node indicates a meeting participant, and its color indicates her hand labeled FE category; the same color scheme is used as in Fig.6; The edge between persons and bar chart on the edge indicate the distribution of emotional interaction between them. Edges are drawn in the same color scheme as in Fig.5. Left red, middle green, and right blue bars in each bar chart indicate the frequency or posterior distribution of empathy, unconcern, and antipathy, respectively.

TABLE I

ESTIMATION PERFORMANCE OF THE PROPOSED METHOD

(a) Estimation performance of emotional interaction given by gaze and FE (Cond. A)

BC	Majority-based agreement rates			
Total	Total	Empathy	Unconcern	Antipathy
0.874	0.673	0.729	0.628	0.671

BC: Bhattacharyya coefficient

(b) Estimation performance of emotional interaction in the joint estimation of emotional interaction and FE given by gaze (Cond. B)

BC	Majority-based agreement rates			
Total	Total	Empathy	Unconcern	Antipathy
0.873	0.675	0.750	0.616	0.004

(c) Recognition performance (precision) of facial expression in the joint estimation of emotional interaction and FE given by gaze (Cond. B)

Method	Total	Nt.	Sm.	Lg.	Wr.	Th.	Ot.
JntEst	0.684	0.662	0.703	0.438	0.000	0.000	0.085
SVM	0.668	0.650	0.704	0.373	0.000	0.000	0.095

JntEst: joint estimation of emotional interaction and FE,
SVM: using only the SVM-based FE recognizer

was ambiguous. Moreover, the corresponding hand labels of facial expression are shown in Fig.6.

E. Comparison with human labels

To quantitatively evaluate the estimation performance of the proposed method, this paper introduces two performance measures that can indicate the degree of matching between the distribution of the hand labels and estimated posterior distributions $P(e|Z_{1:T})$. One measure is Bhattacharyya coefficient[3] between probability distributions p and q, $\mathrm{BC}(p,q) = \sum_e \sqrt{p(e)q(e)}$ $(0 \leq \mathrm{BC} \leq 1)$. The hand labels were first merged into "empathy" $(+2,+1)$, "unconcern" (0), or "antipathy" $(-2,-1)$. Then, their frequency distributions were normalized so as to yield a summation of one. The other measure is majority-based agreement rate in discretized states of the distributions between MAP estimate $\hat{e}$ and majority class of hand labels. If the majority class(s) of the hand labels contained $\hat{e}$, the estimate was considered to be correct. For facial expressions, if the MAP estimate of facial expression, $\hat{f}$, matched the hand label, the estimate was considered to be successful.

Table I shows the performance of the proposed method. As shown in Table I (a), when gaze and FE were given (Cond. A), the emotional interactions were estimated with the fairly high Bhattacharyya coefficient of 0.874 and majority-based agreement rates of 0.673. This suggests that if reliable gaze and FE states are given, the proposed method can accurately replicate the variation in human judgment as regards the type of emotional interaction between humans as also demonstrated in VI-D.

Table I (b) shows the BC and majority-based agreement rates obtained by the joint estimation of emotional interaction and FE (Cond. B). They are comparable to those for Cond. A except for the majority-based agreement rates of antipathy. The recognition rates of FEs obtained by the joint estimation, as well as those obtained by using only the FE recognizer, are shown in Table I (c). The recognition rates of neutral expression and smile are satisfactory, while the FE recognizer failed to correctly recognize almost all wry smiles, thinking,

[3]For example, BC is 0.874, when the voting result of hand labels and posterior distribution are $\{3,1,1\}$ and $\{0.548, 0.452, 0\}$, respectively.

and others expressions. The main reasons seem to be that the facial motions involved in these FEs were too small to be reliably detected, and the appearance frequencies of these FEs in our data were too low to sufficiently train the SVMs. Although our FE recognizer could be improved by fine tuning and refining the data, this problem is still challenging even with state-of-the-art FE recognition techniques. This problem needs continuous effort, and the development of successful technique should allow the framework proposed in this paper to yield more accurate estimation of emotional interaction. Furthermore, there is a possibility that our joint estimation offers better FE recognition thanks to contextual information and prior knowledge of the meeting.

VII. CONCLUSIONS AND DISCUSSIONS

This paper presented a novel research framework for the estimation of emotional interactions produced between meeting participants. We defined emotional interaction as a brief contiguous event wherein a pair of people exchange emotional messages, including empathy, antipathy and unconcern, via verbal and non-verbal behaviors, i.e. "*who is sharing empathy/antipathy with whom*". As the key behaviors, we focused on facial expression and gaze, because their combination realizes the rapid and directed transmission of a large number of emotional messages. We built a probabilistic model that represents a hierarchical structure involving the emotional interactions, facial expressions and other behaviors including utterance and gaze direction. Based on the proposed model, emotional interaction were estimated from interpersonal gaze directions, facial expressions, and utterances, by utilizing the Gibbs sampler. An experiment on several four-party conversations demonstrated the promising performance of the proposed method on the estimation of emotional interactions.

In the next step, we would like to evaluate more the proposed framework both qualitatively, e.g. by questionnaire, and quantitatively using a variety of data, i.e. different numbers of participants and conversational types such as cooperative/hostile discussions. We also intend to extend the proposed model by incorporating other non-verbal behaviors such as gesture, posture and vocal expression, with a consideration of the direction and dynamics of emotional interaction, as well as other important higher-level conversation states. Again, the authors believe that automatically discovering interpersonal emotions, which evolve 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.

REFERENCES

[1] 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.

[2] L. Chen and M. P. Harper. Multimodal floor control shift detection. In *Proc. Int'l Conf. Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009)*, 2009.

[3] 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.

[4] T. Chartrand and J. Bargh. The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999.

[5] J. B. Bavelas, A. Black, C. R. Lemery, and J. Mullett. "I show how you feel": Motor mimicry as a communicative act. *J. Personality and Social Psychology*, 50:322–329, 1986.

[6] N. Chovil. Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163–194, 1991.

[7] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[8] M. Mehu and R. I. M. Dunbar. Naturalistic observations of smiling and laughter in human group interactions. *Behaviour*, 145, 2008.

[9] R. E. Kraut and R. E. Johnston. Social and emotional messages of smiling: An ethological approach. *J. Personality and Social Psychology*, 37(9):1539–1553, 1979.

[10] 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.

[11] M. Argyle and J. Dean. Eye contact, distance and affiliation. *Sociometry*, 28:289–304, 1965.

[12] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[13] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6(1):721–741, 1984.

[14] 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 Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[15] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[16] Y. L. Tian, T. Kanade, and J. Cohn. Facial expression analysis. In *Handbook of face recognition*. Springer, 2005.

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

[18] 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 *In Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (FG 2004)*, pages 129–138, 2004.

[19] 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.

[20] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *In Proc. Int'l Conf. Multimodal Interfaces (ICMI 2007)*, pages 38–45, 2007.

[21] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato. Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings. In *Proc. Int'l Conf. Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009)*, 2009.

[22] K. Otsuka, Hiroshi Sawada, and J. Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations. In *In Proc. Int'l Conf. Multimodal Interfaces (ICMI 2007)*, pages 255–262, 2007.

[23] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. In *Proc. the National Academy of Sciences (PNAS)*, volume 106, pages 15274–15278, 2009.

[24] L. Ding and A. Yilmaz. Learning relations among movie characters: A social network perspective. In *In Proc. European Conference on Computer Vision (ECCV2010)*, volume IV, pages 410–423, 2010.

[25] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Ltd., 1994.

[26] 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 *In Proc. Int'l Conf. Multimodal Interfaces (ICMI 2008)*, pages 257–264, 2008.

[27] S. Gorga and K. Otsuka. Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection. In *Int'l Conf. Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*, 2010.

[28] M. Fujimoto, K. Ishizuka, and T. Nakatani. A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme. In *In Proc. Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP2008)*, pages 4441–4444, 2008.

[29] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expression. *J. Image and Vision Computing*, 18(11):881–905, 2000.

[30] R. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *In Proc. IEEE Int'l W. Real Time Computer Vision for Human Computer Interaction at CVPR*, pages 181–200, 2004.