In Proc. 1st International Workshop on Conversational Informatics, pp. 7–12, Kitakyushu, June, 2005

Probabilistic Inference of Gaze Patterns and Structure of Multiparty Conversations from Head Directions and Utterances

Kazuhiro Otsuka[1,3], Yoshinao Takemae[2], Junji Yamato[1], and Hiroshi Murase[3]

[1] NTT Communication Science Laboratories., NTT Corp., Atsugi, 243-0198, JAPAN
[2] NTT Cyber Solution Laboratories., NTT Corp., Yokosuka, 239-0847, JAPAN
[3] Graduate School of Information Science, Nagoya University, Nagoya, 464-8601, JAPAN

Abstract. A novel probabilistic framework is proposed for inferring gaze patterns and the structure of conversation in face-to-face multiparty communication, based on head directions and the presence/absence of utterances of participants. First, we define three classes of conversational regimes, which are characterized by the topology of the gaze pattern; we assume that they indicate the structure of the conversation, i.e. who is talking to whom. Next, the problem is formulated as joint estimation of both regime state from the gaze pattern and utterance, and the gaze pattern from head directions. We then devise a dynamic Bayesian network, called the Markov-switching model. The regime changes over time are based on Markov transitions, and controls the dynamics of the gaze patterns and utterances. Furthermore, Bayesian estimation of regime, gaze pattern, and model parameters are implemented using a Markov chain Monte Carlo method. Experiments on four-person conversations confirm accurate gaze estimation and the effectiveness of the framework toward identification of the conversation structures.

1 Introduction

Face-to-face conversation is one of the most basic forms of communication in our life and is used for conveying/sharing information, understanding others' intention/emotion, and making decisions. To enhance our communication capability beyond conversations on the spot, intense research efforts are being made to enable teleconferencing, archiving/summarizing meetings, and computer-mediated communication associated with social agents/robots. To achieve such prospective applications, the automatic recognition of conversational scenes, which involve interactive human behavior both physically and psychologically, is a basic technical goal. Our study aims to develop a novel framework for analyzing and understanding multiparty face-to-face conversation by modeling the relationship between the structure of the conversation and the nonverbal behavior that appear in it.

Automatic meeting analysis is an emerging research area, and several methods for the recognition of group actions in meeting have been proposed [1, 2]. However, so far, relatively little attention has been paid to the basic structure of conversations, known as participation roles (speaker, addressees, side-participants, etc.) [3], i.e. who is talking to whom. The identification of participation roles is a particularly important function for services such in automatic video summarization/editing and the social-participation robots that are expected. In the face-to-face setting, it has been suggested that the nonverbal behavior play important roles in the conversation, although verbal information is essential. Among various nonverbal behavior, it is widely acknowledged that gaze serves several important functions such as monitoring others, expressing one's attitudes/intentions, and regulating conversation flow [4, 5]. Based on these psychological findings, it is suggested that since people use gaze behavior as an important cue for understanding the participants' roles in a conversation, it should be possible to automatically determine roles by analyzing people's gaze [6, 7].

To analyze gaze behavior during conversations precisely and quantitatively, it is necessary to realize the automatic measurement of gaze direction in a manner that does not interfere with natural conversation. Unfortunately, the current level of eye tracking techniques fails to meet such requirements, despite recent progress [8, 9]. Instead, an approach that substitutes head direction for eye direction has been proposed [10, 11], since recent face tracking techniques make it easier to measure head direction than gaze [12]. This approach is based on the theory that a person tends to focus his/her attention on the person of interest by centering the person in his/her visual field, which results in rotation of head and/or torso, depending on the positions of other participants.

Our study unifies the above two aspects, i)the link between the structure of conversations and nonverbal behavior, and ii)gaze direction can be approximated by head direction, and formulates a framework for simultaneously solving two problems: inferring the structure of conversations from gaze pattern and utterance, and identifying gaze patterns from ambiguous head-direction measurements. To that end, first, we define three classes of conversational regimes, which can be characterized by the topology of the gaze pattern, and are assumed to indicate the structure of conversations. Next, the problem is formulated using the dynamic Bayesian network called the Markov-switching model [13]. The regime state changes over times based on Markovian transition properties, and it controls the dynamics of utterance patterns and gaze patterns, which stochastically yield head-direction measurements. Furthermore, a Bayesian estimation of the joint posterior distribution of all unknowns consisting of regime states, gaze patterns, and model parameters is implemented with the Markov chain Monte Carlo method, called the Gibbs sampler [14]. Experiments using 4-person conversation were conducted to confirm the effectiveness of the method. So far, a hidden Markov model (HMM) and its derivatives like coupled-HMM [15] have been developed for the recognition of human interaction. However, in contrast to these models, which mainly focus on direct causal relationship between visible human actions, our study tries to explore another aspect that hypothesizes a high-level process that governs how people interact within a social context.

This paper is organized as follows. Section 2 defines the conversational regimes. Section 3 proposes the model and estimation algorithm. Section 4 shows experimental results. Finally, some discussion and our conclusion are presented in Section 5.

2 Conversation Structures and Gaze Patterns

This study aims to develop a framework for the automatic estimation of the structure of multiparty conversation from nonverbal behavior, which can be extracted from audio and visual information. As the structure of conversation, we target *participation role* such as speaker, addressees, and side-participants [3], i.e. who is talking to whom, and who is listening to whom, and the dynamics of how the structure changes over time. To that end, we hypothesize that the stream of a conversation can be segmented into a series of short periods, we call *regimes*, which satisfy two conditions: i) a specific type of nonverbal behavior is continuously present during the regime, and ii)each regime corresponds to a kind of conversation structure, and its temporal changes represent the dynamics of conversations. If such regimes could be extracted and well-defined, the structure of a conversation could be identified by observing and analyzing the sequence of nonverbal behavior. As the nonverbal behavior, we focused on the gaze patterns of participants, and found that there exist a typical topology of gaze patterns during conversations, which frequently appear and have larger temporal scales than individual gaze directions. Moreover, our experimental results suggest a strong link between gaze topology and the conversational structures such that gaze-based video editing can facilitate the viewer's understanding of recorded conversations [16]. Based on these observations, this paper hypothesizes three categories of conversation regimes according to the topologies of gaze patterns: convergence, dyad link, and divergence.

First, the regime called "convergence" corresponds to the gaze pattern in which the gazes from participants converge to one person, i.e. there is one person attracting the others' gazes more than the others, as illustrated in Figure 1(a). This regime corresponds to the conversation structure that one person talks to the others and they look at and listen to the speaker, where the person in center of gaze convergence is the speaker, and the others are the addressees. Here, we denote the regime as R_i^C, where i indicates the center person. This regime is related to past findings such "people gaze more while listening than while speaking"[5].

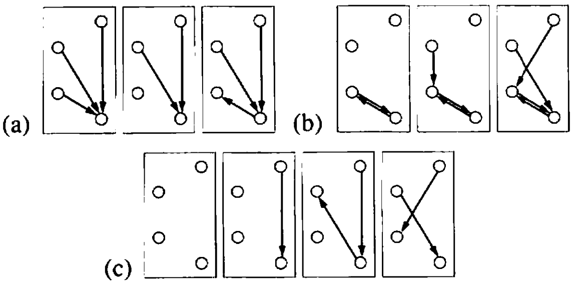
Second, the regime called "dyad link" corresponds to the situation that two people look at each other, i.e. mutual gaze, as illustrated in Figure 1(b). During the regime, they exchange messages and could swap their roles of speaker and addressee; the others are side-participants. This regime often appears during turn taking/giving, and is related to findings that "speakers ended an utterance with prolonged gaze to indicate that it was the turn of one listener to speak" [4, 17]. This regime is denoted as $R_{(i,j)}^{DL}$, where (i,j) represents the pair forming the dyad link.

Third, the regime called "divergence" corresponds to the gaze patterns that do not match the above two regimes, i.e. people look in different directions, as shown in Figure 1(c). In this regime, group conversation does not occur. This often occurs before a conversation starts or at a break point between topics. This regime is denoted as R^0.

3 Model and Estimation

3.1 Notations

This study targets N-person face-to-face conversations($N \geq 3$). The participants are separately seated in chairs, and no one leaves/enters during the conversation. No tools such as notes or whiteboards are used so as not to disturb the attention of the participants. Gaze direction was discretized to N exclusive states: look at the face of one of the other participants or avert from all of them. Let $X_{i,t}$ be the gaze state of



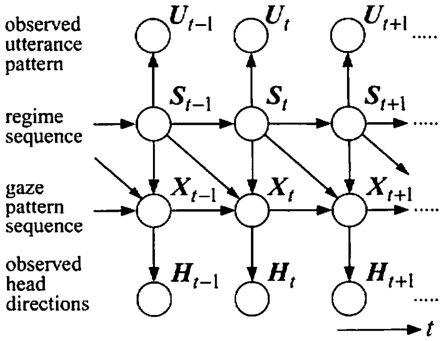


Fig. 1. (*Left*)Typical gaze patterns in each regime: (a)convergence, (b)dyad link, (c)divergence, in the case of 4-person conversation, (node: person, edge: gaze direction, node without outgoing edge: averted gaze).

Fig. 2. (*Right*)Graph representation of structure of Markov-switching model.

person i; looking at person j if $X_{i,t} = j, (i \neq j)$ or avert if $X_{i,t} = i$, at time step t. We call the set of gaze states of all participants the gaze pattern, $\boldsymbol{X}_t = \{X_{1,t}, X_{2,t}, \cdots, X_{N,t}\}$, which takes one of N^N possible patterns. The sequence of gaze pattern is denoted by $\boldsymbol{X}_{1:T} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_T\}$. Let S_t be the regime at time t; it is one of M-regimes as $S_t \in \boldsymbol{R} = \{R^C_i | i = 1 \cdots, N\} \cup \{R^{DL}_{(i,j)} | i = 1, \cdots, N, j = i, \cdots, N, i \leq j\} \cup R^0$, where $M = N + {}_NC_2 + 1$. The sequence of regimes is represented as $\boldsymbol{S}_{1:T} = \{S_1, S_2, \cdots, S_T\}$.

At each time step t, the head direction $h_{i,t}$ of each person i is observed as azimuth angle between world coordinate X and frontal direction of face, as shown in Figure 3(a). We denote the sequence of observed head directions as $\boldsymbol{H}_{1:T} = \{H_1, \cdots, H_T\}, H_t = \{h_{1,t}, \cdots, h_{N,t}\}$. Also, state of utterance is denoted by $u_{i,t} = 1$ if person i utters and $u_{i,t} = 0$ if person i is silent, at time t; the resulting sequence is represented as $\boldsymbol{U}_{1:T} = \{U_1, \cdots, U_T\}$, $U_t = \{u_{1,t}, \cdots, u_{N,t}\}$.

3.2 Model structure

To model the relationship between variables and their temporal evolution, a class of dynamic Bayesian network called the Markov-switching model is defined as shown in Figure 2. In Figure 2, nodes represent variables and edges represents dependencies between variables. This model includes regime sequence $\boldsymbol{S}_{1:T}$ and the sequence of gaze patterns $\boldsymbol{X}_{1:T}$; both of them are considered to be hidden random variables. Here, regime dynamics is assumed to be a first order Markov process with initial probability $P(S_0 = R) = \pi_{0,R}, R \in \boldsymbol{R}$ and transition probability $P(S_t = R' | S_{t-1} = R) = \pi_{R,R'}$, which are constant over time. These model parameters are denoted as $\boldsymbol{\pi}_R = \{\pi_{R,R'} | R' \in \boldsymbol{R}\}$, $\boldsymbol{\Pi} = \boldsymbol{\pi}_0 \cup \{\boldsymbol{\pi}_R | R \in \boldsymbol{R}\}$. The sequence of gaze patterns $\boldsymbol{X}_{1:T}$ are stochastically generated and evolved depending on the emission/transition probabilities $P(\boldsymbol{X}_t | S_t)$,$P(\boldsymbol{X}_t | \boldsymbol{X}_{t-1}, S_{t-1})$, conditioned on the regime state. The likelihood that a gaze pattern $\boldsymbol{X}_t$ appears is given as the product of emission probability $P(\boldsymbol{X}_t | S_t)$ and transition probability $P(\boldsymbol{X}_t | \boldsymbol{X}_{t-1}, S_{t-1})$, as written in

$$P(\boldsymbol{X}_t | \boldsymbol{X}_{t-1}, S_t, S_{t-1}) = P(\boldsymbol{X}_t | S_t) P(\boldsymbol{X}_t | \boldsymbol{X}_{t-1}, S_{t-1}) \tag{1}$$

$$P(\boldsymbol{X}_t | S_t) = \textstyle\prod_{i=1}^{N} P(X_{i,t} | S_t), \quad P(\boldsymbol{X}_t | \boldsymbol{X}_{t-1}, S_{t-1}) = \textstyle\prod_{i=1}^{N} P(X_{i,t} | X_{i,t-1}, S_{t-1}), \tag{2}$$

where we assume the conditional independency of gaze directions of each person for a given regime state. Here, we denote gaze-related model parameters as $P(X_{i,t} = j | S_t = R) = \theta_{R,i,0,j}$, $P(X_{i,t} = j | X_{i,t-1} = k, S_{t-1} = R) = \theta_{R,i,k,j}$, $\boldsymbol{\theta}_{R,i} = \{\theta_{R,i,k,j} | k = 0, \cdots, N; j = 1, \cdots, N\}$, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{R,i} | R \in \boldsymbol{R}, i = 1, \cdots, N\}$. Also, the model includes observation processes that stochastically yield both head directions with probability $P(H_t | X_t)$ conditional to the gaze patterns, and utterance patterns with $P(U_t | S_t)$ for given regime state, at each time step t. We assume all observations at each time step are independent, and also head directions and utterances are independent. Likelihood function of head direction H_t for given gaze pattern $\boldsymbol{X}_t$ is defined using Gaussian distribution so as to reflect uncertainty in head direction, as written in

$$P(H_t | \boldsymbol{X}_t) = \textstyle\prod_{i=1}^{N} p(h_i | X_i), \quad p(h_i | X_i = j) = (2\pi\sigma^2_{i,j})^{-1/2} \exp\left[-(\mu_{i,j} - h_i)^2 / (2\sigma^2_{i,j})\right], \tag{3}$$

where $\mu_{i,j}$, $\sigma^2_{i,j}$ are the mean and variance of the likelihood distribution when person i looks at j, respectively. Also, the independency of head directions of each person for a given gaze pattern, and the temporal invariance of these parameters are assumed. Also, the likelihood of utterance pattern is defined as $P(U_t|S_t) = \prod_{i=1}^{N} P(u_{i,t}|S_t)$, where we assume the utterance of each person at a time step occurs independently but conditional on regime state, and is generated by a Bernoulli process with utterance probability $P(u_{i,t} = 1|S_t = R) = \eta_{R,i}$.

3.3 Bayesian estimation via Gibbs sampling

Based on the model proposed, the problem is to estimate the regime sequence $\boldsymbol{S}_{1:T}$, gaze pattern sequence $\boldsymbol{X}_{1:T}$, and model parameters $\varphi = \{\boldsymbol{\Pi}, \boldsymbol{\Theta}, \{\mu_{i,j}\}_{i,j}, \{\sigma^2_{i,j}\}_{i,j}, \{\eta_{R,i}\}_{R,i}\}$, from measurements $\boldsymbol{Z}_{1:T} = \{\boldsymbol{H}_{1:T}, \boldsymbol{U}_{1:T}\}$. We employ a Bayesian approach to estimate the joint posterior distribution $p(\boldsymbol{S}_{1:T}, \boldsymbol{X}_{1:T}, \varphi|\boldsymbol{Z}_{1:T})$ of all unknown variables for given measurements. In Bayesian analysis, a priori knowledge about the model is represented as the prior distributions of model parameters. To estimate the joint posterior, this study uses a Markov chain Monte Carlo method called the Gibbs sampler [14], which has an advantage when 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.

This study employs natural conjugate prior distributions [18]. Dirichlet distributions are used for the initial and transition probabilities of the regime state, and for emission/transition probabilities of gaze pattern. Priors for head-direction employ Gaussian distributions and inverse chi-squared distributions for mean and variance of its likelihood, respectively. Beta distribution is used for priors of utterance probabilities. Also, full conditional posterior distributions of each variable have the same function form as its priors. Gibbs sampling iterates a set of procedures $\mathcal{N}$ times, and in each step, each variable is sequentially replaced by a new value that is sampled from its full conditional. For example, the regime state S_t and the gaze pattern $\boldsymbol{X}_t$ are sampled from their full conditionals, respectively, as written in

$$P(S_t = R|\boldsymbol{S}_{1:T} \setminus S_t, \boldsymbol{X}_{1:T}, \varphi, \boldsymbol{Z}_{1:T}) \propto P(S_t = R|S_{t-1})P(X_t|S_t = R)P(U_t|S_t = R)P(S_{t+1}|S_t = R), \quad (4)$$

$$P(\boldsymbol{X}_t|\boldsymbol{S}_{1:T}, \boldsymbol{X}_{1:T} \setminus \boldsymbol{X}_t, \varphi, \boldsymbol{Z}_{1:T}) \propto P(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}, S_{t-1})P(\boldsymbol{X}_t|S_t)P(H_t|\boldsymbol{X}_t)P(\boldsymbol{X}_{t+1}|\boldsymbol{X}_t, S_t). \quad (5)$$

After the iterations terminate, statistics are calculated from the samples $\{\boldsymbol{S}^{(q)}_{1:T}, \boldsymbol{X}^{(q)}_{1:T}, \varphi^{(q)}\}$ for iteration steps $q = \mathcal{N}'$ to $\mathcal{N}$ to ensure convergence. For regime sequence and gaze sequence, the maximum a posterior estimate is calculated as $\hat{S}_t = \arg\max_{R \in \boldsymbol{R}} \sum_{q=\mathcal{N}'}^{\mathcal{N}} \delta_R(S^{(q)}_t)$, where $\delta_R(R') = 1$ if $R = R'$, otherwise $\delta_R(R') = 0$. For other variables, the minimum mean-squared error estimates are calculated as in $\hat{\mu} = (\mathcal{N} - \mathcal{N}' + 1)^{-1} \sum_{q=\mathcal{N}'}^{\mathcal{N}} \mu^{(q)}$.

4 Experiment

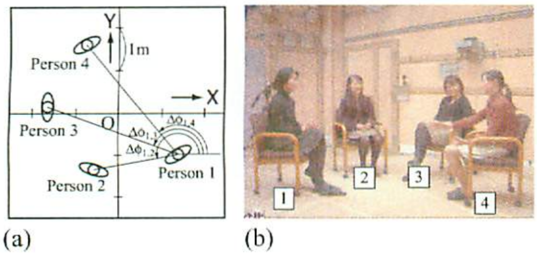
4.1 Recording data and initial setting

Data were recorded for 4-person group conversations. The participants were four women within the same age bracket. They were instructed to have a discussion and try to reach a conclusion as a group for a given discussion topic ("*Is marriage and love same or different?*"), within five minutes. The head directions were measured at 30Hz with magnetic-based 6-DOF sensors (POLHEMUS Fastrak$^{\mathrm{TM}}$), which were attached to their heads with hair bands. Figure 4(a) shows the first 3600 time steps (=2[min]) of head azimuth of each participant. Audio data were recorded with clip-on microphones attached to each participant, and utterance intervals were manually extracted using a waveform editor. Figure 4(b) shows the utterance intervals of each participant. Also, video sequences, whole shot (Figure 3(b)) and bust shots (Figure 7(a)), were recorded at 30[frame/sec]. These data were synchronized and 10000 time steps (=frames, $\simeq$ 5.6[min]) were used in the analysis. Ground truth of gaze direction at each time step was manually created by watching the video sequences.

Hyper parameters for prior distribution were set based on the following policy. The bearing angles $\Delta\phi_{i,j}$ given by the relative positions of participants, were employed as the mean values of prior distribution of head-direction likelihood (See Figure 3(a)). In regime 'convergence', the gaze-direction distribution of the speaker is set to uniform, while others look at the speaker with high probability (0.7). In regime 'dyad link', the pair look at each other with high probability (0.95), while the two others look around randomly. In regime 'divergence', people look at various directions with uniform probability.

4.2 Results

Estimation results were obtained after $\mathcal{N} = 700$ iterations of Gibbs sampling ($\mathcal{N}' = 500$). Figure 5(a), which shows the transition of the mean $\{\mu_{1,j}\}_{j=1}^{4}$ of head-direction likelihood distribution as a function



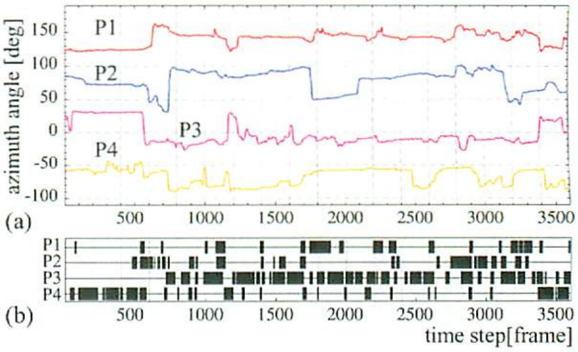


Fig. 3. (*Left*)Overview of scene. (a)plan view of participants' allocation, (b)whole view of participants.

Fig. 4. (*Right*)Observed data for 2[min], (a)head azimuth, (b)temporal intervals with utterance, for participants.

of iteration step number, shows that convergence was achieved. Figure 6(a) shows the estimation result of gaze direction and the corresponding ground truth, illustrated for a 2[min] period. Average correct ratio of the number of frames wherein estimates and ground truth coincide, was 71.1%. Most errors were related to the 'avert' gaze status. This is because human can avert/turn their gaze from/on someone without moving their head, e.g. using a sidelong glance. Also, the cause of the error can be explained by Figure 5(b), which shows the estimated distributions of head-direction likelihood and histograms of head direction for separate gaze directions. In Figure 5(b), both distributions exhibit significant overlaps between that for averted gaze and those of the others. In addition, the average correct ratio of maximum a posteriori estimates based on the ground truth of gaze direction was 68.8%. Given that our result from 'unsupervised learning' was better than one from 'supervised learning', it is suggested that the proposed framework is an effective methodology for detecting gaze direction.

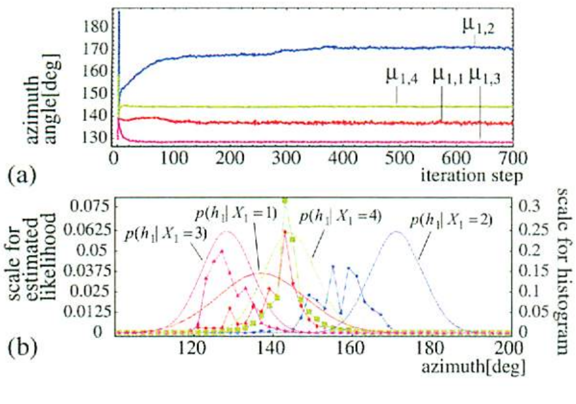
Figure 7 shows an example of regime transition; $R^C_4 \to R^C_{(2,4)} \to R^C_2$. Figure 7(a) shows bust-shot images of each participant and Figure 7(b) shows gaze patterns. At first ($t = 310$), person 4 talked to all others (P4: *"Even if I am not thinking of marriage, I have to think about having relations, I mean.."*) and others listen to person 4. This form of conversation was indicated by estimated regime R^C_4. Next ($t = 485$), person 2 responded to person 4 saying (P2: *"Yes, yes, yes, yes, yes"*) with nodding, and P4 looked at P2 to confirm the response from her. There was mutual gaze between person 2 and 4, which is indicated by regime estimate, dyad link $R^{DL}_{(2,4)}$. Furthermore ($t = 578$), P2 keep on speaking (P2: *"yes, in terms of ever since"*) and person 4 returned response back to P2 saying (P4: *"yes, yes"*) and then stopped speaking, which indicated that P4 was offering the floor to P2. At the same time, person 3 turned her gaze from P4 to P2, in order to watch what P2 would say. From the above results, it is confirmed that the estimated regimes seem reasonable and could be used as an indicator of conversation structure.

5 Discussion and Conclusion

A probabilistic framework based on head directions and utterances was proposed for inferring gaze patterns and the structure of conversations in face-to-face multiparty communications. To that end, we devised the Markov-switching model, whose hidden states correspond to the regime and gaze patterns. A Bayesian estimation of all unknown variables including model parameters is carried out using the Gibbs sampler. Experiments on four-person conversations confirmed the effectiveness of our framework. As the next step, it is necessary to evaluate the sequence of regime estimates by comparing them with actual events that take place during conversations. Also, we need to increase the amount of data so that it includes different people, different group size, and various actions. such as locomotion and note-taking. The proposed framework can be extended to incorporate other human behavior such as head gestures like nodding and shaking, facial expressions, and prosody. Furthermore, real-time online estimation and image-based head tracking are required to develop practical applications.

References

1. I. McCowan et al., "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. PAMI*, Vol. 27, No. 3, 2005.
2. D. Zhang et al. "Modeling individual and group actions in meetings: a two-layers HMM framework," *Proc. 2nd IEEE Workshop on Event Mining*, 2004.



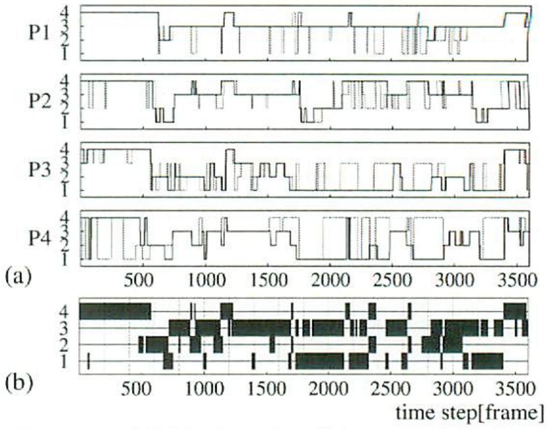


Fig. 5. (*Left*) (a)transition of $\mu_{1,1}, \mu_{1,2}, \mu_{1,3}, \mu_{1,4}$ through iteration of Gibbs sampler, (b)estimated likelihood function $p(h_1|X_1 = i)$ (person 1 look at person if $i \neq 1$, or avert gaze if $i = 1$), and line with symbol shows corresponding histogram, (symbol = diamond: avert, star: gaze at P2, triangle: gaze at P3, square: gaze at P4).

Fig. 6. (*Right*)Estimated sequences of (a)gaze pattern $\{X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}\}$ and (b)regime states. In (a), solid lines : estimates, dashed lines : ground truth. In (b), single band at a time slice indicates regime R_i^C (convergence), dual band at time slice indicates regime $R_{(i,j)}^{DL}$ (dyad link), and no band indicates R^0 (divergence).

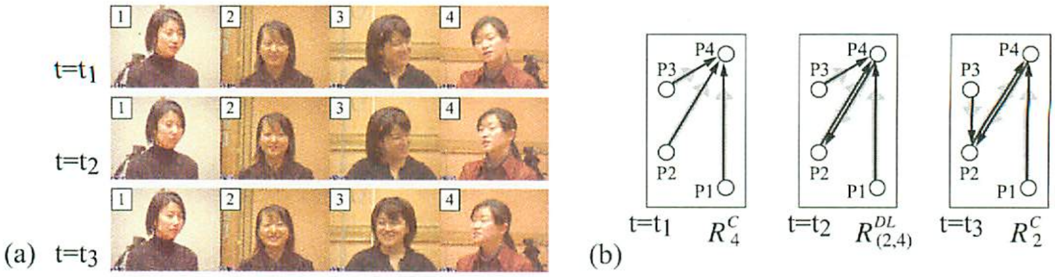


Fig. 7. An example of regime transition ($t_1 = 310, t_2 = 485, t_3 = 578$). (a)snapshot of each participant, (b)regime estimates and gaze patterns (solid arrows: estimates, wide arrows: ground truth).

3. H. H. Clark and T. B. Carlson, "Hearers and speech acts," *Language*, Vol. 58, pp.332–373, 1982.
4. A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, Vol. 26, pp. 22–63, 1967.
5. M. Argyle and M. Cook, *Gaze and Mutual Gaze*, Cambridge Univ. Press, 1976.
6. N. Jovanovic and R. Akker, "Towards automatic addressee identification in multi-party dialogues," *Proc. SIGdial*, pp. 89–92, 2004.
7. Y. Takemae, K. Otsuka, and N. Mukawa, "An analysis of speakers' gaze behavior for automatic addressee identification in multiparty conversation and its application to video editing," *Proc. IEEE RO-MAN*, pp. 581–586, 2004.
8. T. Ohno and N. Mukawa, "A free-head, simple calibration, gaze tracking system that enables gaze-based interaction," *Proc. Eye Tracking Research & Application Symposium (ETRA) 2004*, pp. 115–122, 2004.
9. Y. Matsumoto and A. Zelinsky, "An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement," *Proc. Int. Conf. Automatic Face and Gesture Recognition*, pp. 499–504, 2000.
10. R. Stiefelhagen et al., "Modeling focus of attention for meeting index based on multiple cues," *IEEE Trans. Neural Networks*, Vol. 13, No. 4, 2002.
11. D. Reidsma et al., "Virtual meeting rooms: from observation to simulation," *Proc. Social Intelli. Design*, 2005.
12. L.-P. Morency et al., "Adaptive view-based appearance model," *Proc. CVPR'03*, pp. 803–810, 2003.
13. C.-J. Kim and C. R. Nelson, *State-space models with regime switching*, MIT Press, 1999.
14. W. R. Gilks et al. *Markov chain Monte Carlo in practice*, Chapman & Hall/CRC, 1996.
15. N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. PAMI*, Vol. 22, No. 8, 2000.
16. Y. Takemae, K. Otsuka, and N. Mukawa, "Impact of video editing based on participants' gaze in multiparty conversation," *Proc. CHI2004*, pp. 1333–1336, 2004.
17. D. G. Novic et al. "Coordinating turn-taking with gaze," *Proc. Int. Conf. Spoken Lang.* pp. 1888–1891, 1996.
18. J. M. Bernardo and A. F. M. Smith, *Bayesian theory*, John Wiley & Sons, Ltd., 1994.