# CONVERSATION SCENE ANALYSIS WITH DYNAMIC BAYESIAN NETWORK BASED ON VISUAL HEAD TRACKING

*Kazuhiro Otsuka* [*] *Junji Yamato* *Yoshinao Takemae* *Hiroshi Murase*

NTT Communication Science Labs.
{otsuka,yamato}@eye.brl.ntt.co.jp

NTT Cyber Solutions Labs.
takemae.yoshinao@lab.ntt.co.jp

Nagoya University
murase@is.nagoya-u.ac.jp

## ABSTRACT

A novel method based on a probabilistic model for conversation scene analysis is proposed that can infer conversation structure from video sequences of face-to-face communication. Conversation structure represents the type of conversation such as monologue or dialogue, and can indicate who is talking / listening to whom. This study assumes that the gaze directions of participants provide cues for discerning the conversation structure, and can be identified from head directions. For measuring head directions, the proposed method newly employs a visual head tracker based on Sparse-Template Condensation. The conversation model is built on a dynamic Bayesian network and is used to estimate the conversation structure and gaze directions from observed head directions and utterances. Visual tracking is conventionally thought to be less reliable than contact sensors, but experiments confirm that the proposed method achieves almost comparable performance in estimating gaze directions and conversation structure to a conventional sensor-based method.

## 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 reaching decisions. In recent years, meeting scene analysis has emerged as an attractive way of creating innovative multimedia applications for teleconferencing, archiving / summarizing meetings, and social agents / robots. Several attempts have been made to achieve the automatic recognition of group actions in meetings using HMMs [1], layered-HMM [2], and dynamic Bayesian networks [3, 4]. Current approaches mainly utilize the direct modeling of physical actions to recognize interaction between people, and little attention has been paid to the underlying structure of conversations that rules human interaction.

In contrast to the existing approach, we have been trying to explore the underlying structure of conversations; this structure governs how people interact within the social context of conversation. As a basic conversation structure, we focus on the combination pattern between participants and their participation roles such as speaker, addressees, and side-participants [5]. This conversation structure can indicate who is talking / listening to whom, and is an essential element in describing conversation scenes. To extract this structure, we focus on eye gaze, which is acknowledged to serve important functions such as cues for addressing and listening [6], and controlling turn-yielding/taking [7]. Based on these empirical findings, we have hypothesized that the structural features of gaze pattern among participants can characterize the conversation structure, and have proposed a probabilistic conversation model based on a dynamic Bayesian network that represents the hierarchical relationship between the conversation structure and human behavior [8]. In this model, gaze directions are inferred from head directions, because the direct measurement of gaze direction is difficult. Furthermore, the Markov chain Monte Carlo (MCMC) method [9] is used to realize Bayesian estimation of the conversation structure and gaze pattern from observed head directions and utterances. In [8], head direction was accurately measured by attaching a magnetic-based sensor to each participant, and quantitative evaluations confirmed that the sensor-based head direction was effective for estimating gaze direction and the conversation structure. Although such sensors are accurate, they are not practical and interfere with natural communication.

To support a broad range of applications, this paper proposes a new method for conversation scene analysis that measures head direction by tracking the heads of the participants in video sequences captured by monocular cameras, instead of using contact-type sensors. As the head tracking method, we employ Sparse-Template Condensation (STC) tracker [10]. The sparse template consists of a sparse set of feature points within a rectangle region. The human face is approximated as a planar surface forming the sparse template. Condensation algorithm is used to estimate the posterior density of state of the template that represents the position and direction of the face. STC tracker provides significant robustness against partial occlusions, such as those that occur in profile shots. The measured head directions are input to the conversation model that the authors previously proposed [8]. Experiments confirm that proposed method achieves reasonable performance

[*] First author is also with Nagoya University.

in estimating both gaze direction and conversation structure.

This paper is organized as follows. Section 2 reviews the conversation model and estimation algorithm, and Section 3 presents head tracker. Section 4 shows experiment results that verify the performance of the proposed method. Finally, our conclusion and discussions are presented in Section 5.

## 2. CONVERSATION MODEL AND ESTIMATION

This study targets small group conversations held in a closed environment, and aims to realize the automatic identification of the structure of multiparty conversations from human behavior extracted from audio and visual information by a probabilistic conversation model [8]. The conversation model is based on a dynamic Bayesian network called the Markov-switching model [11]. The Markov-switching model exhibits a hierarchical structure; a discrete random process at a higher level evolves through Markovian transitions, and it governs the dynamics of the processes at the lower levels. Here, the high-level process corresponds to the conversation structure, we call it the conversation regime, and the lower one corresponds to participant behavior which includes gaze and utterance patterns. The temporal changes in regimes are assumed to represent the dynamics of conversation such as turn-taking.

Assuming that gaze patterns can characterize the conversation structure, three classes of conversation regimes are defined based on the structural features of gaze patterns; i) convergence, ii) dyad-link, and iii) divergence. Regime convergence $\boldsymbol{R}^C$ is assumed to correspond to a monologue scene, and is indicated by the convergence of gazes to one person (Fig.1(a)). Regime dyad-link $\boldsymbol{R}^{DL}$ is assumed to correspond to dialogue scenes, and is indicated by mutual gaze between two people (Fig.1(b)). Regime divergence $R^0$ is assumed to represent scenes that do not match the above two regimes, i.e. people look in different directions or avert their gaze (Fig.1(c)).

Fig. 2 shows the structure of the conversation model. Hidden variables include regime state sequence $\boldsymbol{S}_{1:T} = \{S_t\}_{t=1}^T$ and the sequence of gaze patterns $\boldsymbol{X}_{1:T} = \{\boldsymbol{X}_t\}_{t=1}^T$. The regime state at time step $t$, $S_t$, takes one of $M (= N + {}_NC_2 + 1)$-regimes as $S_t = R \in \boldsymbol{R} = \boldsymbol{R}^C \cup \boldsymbol{R}^{DL} \cup R^0$, where $N$ denotes the number of participants. The gaze pattern $\boldsymbol{X}_t$ is composed of the set of gaze directions of all participants, $\boldsymbol{X}_t = \{X_{i,t}\}_{i=1}^N$, where $X_{i,t}$ denotes the gaze state of person $i$; looking at person $j$ if $X_{i,t} = j, (i \neq j)$ or avert if $X_{i,t} = i$. Observable variables $\boldsymbol{Z}_{1:T}$ consist of the sequences of head directions $\boldsymbol{H}_{1:T} = \{\boldsymbol{H}_t\}_{t=1}^T$ and utterance patterns $\boldsymbol{U}_{1:T} = \{\boldsymbol{U}_t\}_{t=1}^T$. The head direction $h_{i,t} \in \boldsymbol{H}_t$ of each person $i$ is observed as the azimuth (horizontal) angle between world coordinate X and the frontal direction of face, as shown in Fig 4(a). Also, the utterance pattern $\boldsymbol{U}_t = \{u_{i,t}\}_{i=1}^N$ indicates whether person $i$ is making utterance ($u_{i,t} = 1$), or not ($u_{i,t} = 0$). Based on the model, the problem is to estimate regime sequence $\boldsymbol{S}_{1:T}$ (the conversation structure), gaze pattern sequence $\boldsymbol{X}_{1:T}$, and model parameters $\varphi$, from observed head directions $\boldsymbol{H}_{1:T}$ and utterances $\boldsymbol{U}_{1:T}$. To yield a Bayesian solution, we use the MCMC method called Gibbs sampler; it outputs 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.

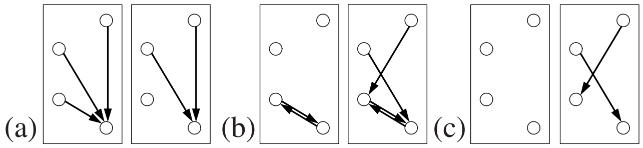

**Fig. 1**. Examples of gaze patterns including typical structures; (a)convergence, (b)dyad-link, (c)divergence. Node: participant. Directed edge : gaze direction, Node without outgoing edge : gaze aversion.

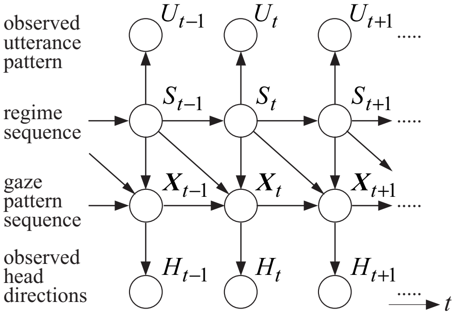

**Fig. 2**. Graph representation of proposed conversation model.

## 3. VISUAL HEAD TRACKING

For measuring head directions, this paper employs the Sparse-Template Condensation (STC) tracker [10]; it is fast and robust. In contrast to traditional template matching, which assesses all pixels in a rectangular region, the sparse template consists of a sparse set of feature points within a rectangular template region. The feature points are located at the local minimum/maximum of image intensities and straddle the zero-cross boundaries of images. Assuming the planar approximation of the human face, this paper manually registers a template that includes facial parts, as shown in Fig. 3(a). The state of a template, which represents the position and pose of the face, is defined as a vector consisting of 2-DOF translation on the image plane, 3-DOF rotation, and scale (we assume weak-perspective projection). The Condensation algorithm is used to sequentially estimate the posterior density of the template state, which is represented as a particle set, as shown in Fig. 3(b). The weight of each particle is calculated based on a robust function for projected feature points on the image. The STC tracker has the advantages of speed owing to the sparseness of the feature points and robustness owing to robust matching, multiple-hypothesis generation/testing by the Condensation approach, and multiple templates which allows partial occlusions to be well handled.

The indeterminate problem of the sign of rotational angle, which is due to the weak-perspective projection, is solved by the following approach; i) impose a depth offset so that the

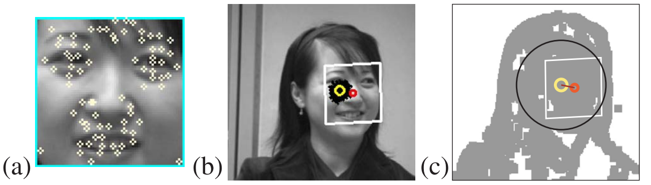

**Fig. 3**. (a)sparse template, (b)template during tracking (white frame); black cloud : particle distribution, larger circle : center point of template (head center), smaller circle : center of template surface. (c)extracted foreground region (gray area) and estimated head circumference (circle).

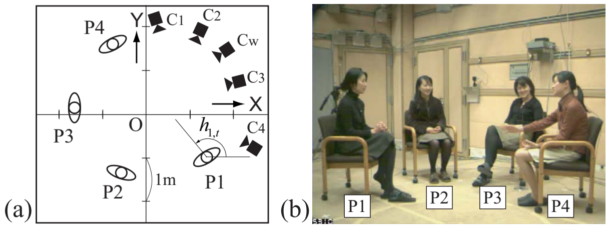

**Fig. 4**. Overview of scene. (a)location of each participant P$i$ and camera C$i$, (b)whole view of participants (G1) from C$_W$.

template plane fits the surface of face and the center indicated by the template state is located at the center of head, ii) penalize a particle if the estimated head circumference overreaches the actual head region extracted as foreground (Fig. 3(c)).

Images of each participant are recorded by separate uncalibrated cameras, as shown in Fig. 4(a) and Fig. 6(a), and his/her head direction is measured in terms of relative angle away from the frontal position toward cameras, as the input of the conversation model. Note our method does not need absolute angle in the world coordinate common to all participants, it requires only relative order in azimuth angles between participants as observed from the participant's position.

## 4. EXPERIMENTS

### 4.1. Data Set

Experiments used the data collected in [8], which includes 4-person group conversations held by two groups G1 and G2; each group consisted of four women, and held two discussions on given topics, we denote them G1-C1, G1-C2, G2-C1, and G2-C2. The videos ranged from 5.1 to 5.6 min. Fig. 4 shows the seat and camera for each participant. The hyper-parameters of the model were also as same as those in [8]. The number of particles used in the STC tracker for each person was set at 2000. Unit time step was 1/30 second. Image size for tracking was 320×240 pixels.

### 4.2. Accuracy of Head Direction

Fig. 5 shows the head azimuth of each participant obtained by STC tracker in the first 3600 time steps (=120 sec.) for G1-C1. Fig. 5 also shows the corresponding sensor output from magnetic-based sensors (POLHEMUS Fastrak™), which were attached to their heads with hair bands. For comparison, a bias was added each STC tracker output so that it had the same average as the sensor output. Fig. 5 indicates that the STC data exhibit significant fluctuation, but approximately replicate the sensor output. The mean deviation in G1-C1 was 5.1, 6.9, 10.0, and 3.7[deg], for subjects $P1 \sim P4$, respectively. For all data, the mean deviation was 5.9[deg]. Fig. 6 shows snapshots of the templates during tracking. Fig. 6(b) shows that self-occlusion, due to the profile view, can successfully be handled. Also, the robustness against changes in facial expression from the one in the registered template is confirmed, as shown in Fig. 6(c). Although partial occlusions sometimes occurred as shown in Fig. 6(d), stable tracking continued through the use of two templates; one is the whole region and the other is the upper-half region for the P3 case. Despite the planar approximation of the face, STC tracker yielded data that was reasonably accurate. The tracking of each person was done offline without termination, while automatic re-initialization was invoked if the track was lost due to full occlusions e.g. face was fully covered by both hands when embarrassed. The processing speed was about 0.20[sec/frame].

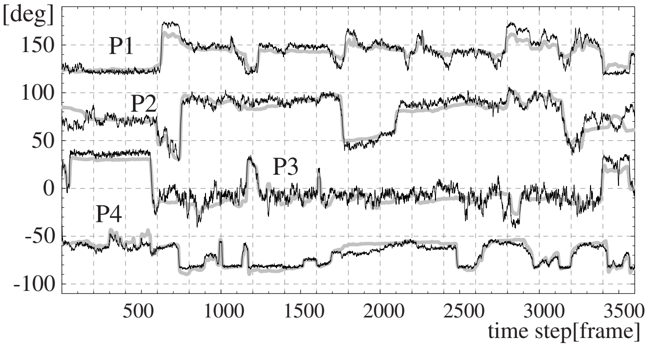

**Fig. 5**. Mesurement data of azimuth head direction. solid line: STC, pale wide line: sensor data.

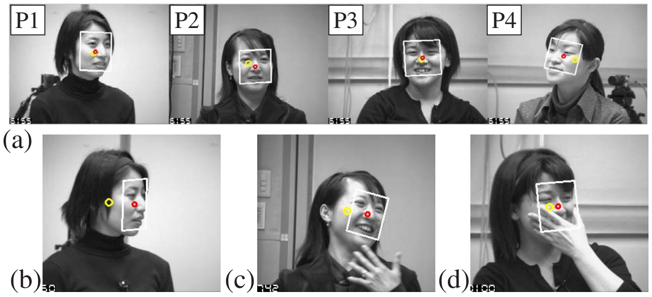

**Fig. 6**. Snapshot of participants during tracking. (a)all participants, (b)profile view, (c)laughing, (d)hand-covering.

### 4.3. Evaluation of Gaze Directions

Fig. 7(a) shows the estimation results of gaze direction and the corresponding ground truth, for the same period shown in Fig. 5. The ground truth of gaze direction was manually created by watching the video sequences. Table 1 shows the

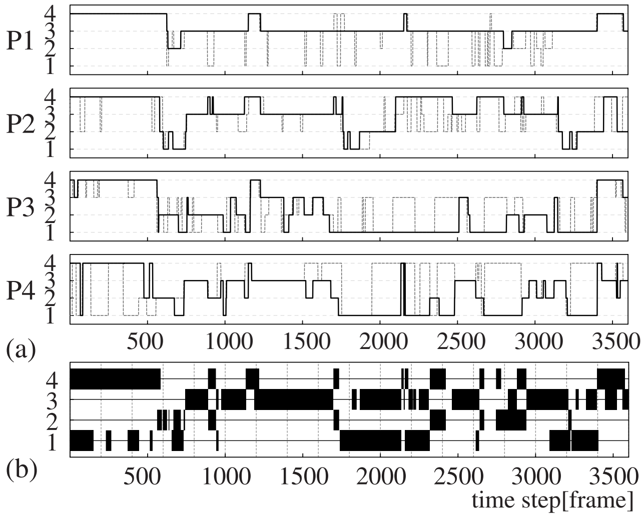

**Fig. 7**. Estimated sequences of (a)gaze pattern $\{X_{i,t}\}_{i=1}^{4}$ and (b)regime states. In (a), solid lines : estimates, dashed lines : ground truth. In (b), single band at a time slice indicates regime $R_i^C$ (convergence), dual band indicates regime $R_{(i,j)}^{DL}$ (dyad link), and no band indicates $R^0$ (divergence).

average correct ratio where estimates and ground truth coincide. A comparison to the results obtained with magnetic-based sensors, [8], which is denoted as "reference" in Table 1, confirmed that the proposed methods can achieve almost comparable accuracy.

### 4.4. Evaluation of Conversation Regimes

Fig. 7(b) shows a part of the estimated regime sequence for G1-C1. Table 2 shows the accuracy of regime estimates for each conversation. The accuracy is defined as the degree of match between the regime estimates and annotations that represent the class and directionality of utterances, given for each utterance interval [8]. Table 2 indicates that the accuracy of the proposed method is reasonably high; it almost matches that of the magnetic-based sensor system.

## 5. CONCLUSION AND DISCUSSION

This paper newly incorporated a visual head-tracking method into a probabilistic conversation model for identifying the conversation structure from audio-visual recordings of meetings. Experiments confirmed that the proposed method offers reasonable accuracy in estimating gaze direction and conversation structure; results show that visual-head tracking is an effective technique for extracting the visual attention of participants. Unlike sensor-based methods, which require special hardware and environments, the visual approach requires only a video camera / recorder, which are commonly available nowadays. The results from this paper show a way to apply the proposed framework of conversation analysis to a wide range of video-based applications such as video archiving and computer-mediated communications.

**Table 1**. Accuracy of gaze direction estimates [%].

| | G1-C1 | G1-C2 | G2-C1 | G2-C2 |
|---|---|---|---|---|
| Proposed | 64.7 | 57.8 | 69.5 | 74.5 |
| Reference | 71.1 | 59.3 | 72.4 | 75.9 |

**Table 2**. Accuracy of regime estimates [%].

| | G1-C1 | G1-C2 | G2-C1 | G2-C2 |
|---|---|---|---|---|
| Proposed | 77.7 | 83.1 | 89.2 | 92.7 |
| Reference | 81.9 | 92.1 | 91.4 | 96.3 |

Future works include the following. First, although the current method needs manual registration of template, face detection techniques such as in [12] can easily be employed to remove this manual process. Also, this paper uses one camera per person, but the head tracker used can be applied to the image frames captured by just one wide-angle camera. For practical meeting situations, it is also important to deal with dynamic scenes including entrance / movement / departure of people during the conversations.

## 6. REFERENCES

[1] I. McCowan, D. Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings,," *IEEE Trans. PAMI*, vol. 27, no. 3, 2005.

[2] D. Zhang, D. G. Perez, S. Bengio, I McCowan, and G. Lathoud, "Modeling individual and group actions in meetings: A two-layers HMM framework," in *Proc. 2nd. IEEE Workshop on Event Mining*, 2004.

[3] A. Dielmann and S. Renals, "Dynamic Bayesian networks for meeting structuring," in *Proc. IEEE ICASSP'04*, 2004.

[4] M. Al-Hames and G. Rigoll, "A multi-modal mixed-state dynamic bayesian network for robust meeting event recognition from disturbed data," in *Proc. ICME'05*, 2005.

[5] E. Goffman, *Forms of Talks*, University of Pennsylvania Press, Philadelphia, 1981.

[6] C. Goodwin, *Conversational Organization : Interaction between Speakers and Hearers*, Academic Press, 1981.

[7] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[8] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase, "A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances," in *Proc. ACM ICMI'05*, pp. 191–198, 2005.

[9] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC, 1996.

[10] Y. Matsubara and T. Shakunaga, "Sparse template matching and its application to real-time object tracking," *IPSJ Trans. CVIM*, vol. 46, no. SIG9, pp. 60–71, 2005, (In Japanese).

[11] C.-J. Kim and C. R. Nelson, *State-Space Models with Regime Switching*, MIT Press, 1999.

[12] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.