Inferring State Transition from Bystander to Participant in Free-style Conversational Interaction

Tatsuya Era, Hiroki Yoshimura, Masashi Nishiyama, and Yoshio Iwai
Graduate School of Engineering, Tottori University, Japan
nishiyama@eecs.tottori-u.ac.jp

Abstract

We propose a novel method for inferring the state transition from bystander to participant in free-style conversational interactions, using physical behaviors acquired from cameras and a microphone. Although existing methods address participants and a presenter, these methods do not consider bystanders, who play an important role in the interaction. In the research field of cognitive science, the existing model considers psychological aspects of changing from bystander to participant. However, this model is difficult to implement because inferring the psychological aspects of bystanders is a challenging task. Instead of using psychological aspects, our method exploits physical behaviors such as standing position, facial direction, and voice direction. We analyzed the suitable parameters of the behavior to increase performance in inferring state transitions, using datasets collected from poster presentations.

1. Introduction

Free-style conversational interaction is an essential form of communication between persons. This interaction does not assume that persons sit down around a table to communicate [3, 10, 11, 14]. We address the assumption that they are instead coming and going as the conversations unfold. It is a very difficult task to infer the state of each person in conversation because each persons movements and speech are continually changing.

The state of each person in free-style conversational interaction can be categorized into the following three types. The first state is a presenter who is making a presentation. The presenter stands near an exhibit and holds the initiative in conversation. The second state is a participant who is speaking to the presenter. The number of participants is dynamic over the course of conversations. The third state is a bystander who is somewhere in the vicinity of the presenter, but does not speak to the presenter. This is the state before a person becomes a participant. The behavior of the bystander has diverse qualities such as specific distance from, and presence or absence of attention to, the presenter and the exhibit. The number of bystanders, like the number of participants, changes continually.

Some researchers [6, 1, 15, 7] have addressed the problem of inferring the state of the presenter and the participants in free-style conversational interaction. The existing methods [6, 1] automatically detect a group that comprises a presenter and participants. These methods [15, 7] use behavioral gestures and voice information to infer the states of the presenter and the participants. However, many existing methods [6, 1, 15, 7] have not focused on the bystander.

The bystander state is also important in free-style conversational interaction. For instance, we can obtain useful information from bystanders such as whether they change into participants by beginning to speak with the presenter, whether they leave the exhibit without speaking to the presenter, whether they stay in the vicinity of a presenter or exhibit for a long time because the presenter has not noticed them, or whether they simply pass by without seeing the presenter or the exhibit. This bystander information has many potential applications, e.g., marketing for conference events such as exhibitions and poster presentations, as well as retail events such as bargain sales. For instance, a presenter can review his or her own presentation by comparing the ratio of bystanders to participants. From a cognitive science viewpoint, Bono et al. [2] have observed state transitions from bystanders to participants. However, this approach is cumbersome because it requires attaching a camera and a microphone to the body of each person and determining the state of each person manually, using visual or auditory information.

To this end, we propose a novel method for automatically inferring transitions from bystander to participant, using non-body-attached cameras and a microphone. To represent the process of this state change from bystander to participant, we designed the state transition model, using the following physical parameters: standing positions and facial direction acquired from the cameras, and voice direction acquired from the microphone. We collected video

sequences of free-style conversational interactions and evaluated the performance of the model for inferring state transitions from bystander to participant.

The rest of this paper is organized as follows. Section 2 describes our state transition model; Section 3 describes the method for determining the state transition; Section 4 presents the results of our experiments and analyses; and our concluding remarks are given in Section 5.

2. State transition model from bystander to participant

2.1. Overview

Existing models for detecting states of persons in free-style conversational interaction [5, 4] use psychological dimensions. Bono et al. [2] expanded these models. However, even the expanded models determine state transitions manually after an annotator has labeled the changes in psychological aspects. Therefore, existing models cannot be implemented easily on a computer. To automate the determination of state transitions, we focus on physical behaviors. Bono et al. [2] noted that changes in psychological aspects of bystanders and participants are observable as physical behaviors such as standing position, facial direction, and voice direction. We consider a poster presentation a concrete example of free-style conversational interaction. We assume that there are multiple bystanders and participants for one presenter, and by definition, bystanders do not converse with anyone, including other participants.

2.2. Preliminary observation of free-style conversational interaction

It is well known that persons who are talking to each other create an F-formation [8]. This formation refers to the phenomenon that persons in conversation are continuously adjusting their standing positions and facial directions with respect to each other. The mutual process of approving someone to join a conversation is performed before one enters the F-formation. McNeill [9] pointed out that not only persons, but also exhibits, are elements of F-formations.

We investigated what kind of behavior was performed during a poster presentation. We observed that when they spoke to the presenter, participants stood within a semicircular area with the exhibit in the center. Figure 1 shows the positions of the participants, the presenter, and the exhibit in F-formation at a poster presentation. The bystanders who were not talking with the presenter stood outside the formation. Note that it sometimes happened that some bystanders stood in the F-formation before the process of approval to join conversation was complete. We term this situation as apparent F-formation.

We interviewed the bystanders who stood in the apparent F-formation, and the presenter, after the presentation. All

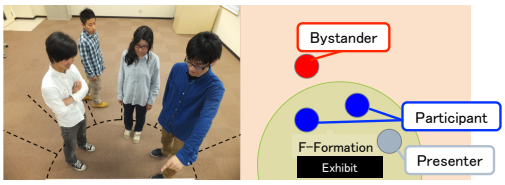

Figure 1. Example of F-formation formed by a bystander, two participants, and a presenter.

of the following situations were reported: The presenter did not notice the presence of the bystander; the bystander did not want to talk to the presenter; the bystander approached to read the contents of the exhibit; and the bystander waited for the opportunity to talk with the presenter. On the basis of our observations and information provided by Bono [2], we designed the state transition model that follows, using these physical behaviors.

2.3. Design of the state transition model

2.3.1 Describing states

In Figure 2 we summarize each state of the bystander and the participant in the model:
S1: standing in the area (the region covered by cameras and microphone),
S2: viewing the exhibit or the presenter from a distance,
S3: not joining the conversation in apparent F-formation,
S4: joining the conversation.
S1 represents the situation in which the bystander, who does not join the conversation, simply stands in the area. The physical aspect of S1 is that he/she does not have the exhibits or presenters in sight. The psychological aspect of S1 is that he/she does not notice the conversation. S2 represents the situation that the bystander views the exhibit or the presenter from a distance. The psychological aspect of S2 is that he/she notices the exhibit or the presenter, and views it to learn what kind of exhibit is displayed. S3 represents the situation in which the bystander stands in an apparent F-formation while not joining the conversation. This situation happens when the presenter has not approved the bystander to join the conversation. The psychological aspect of S3 is that the bystander is interested in the presenter or the exhibit. S4 represents the bystanders joining the conversation as a participant. The psychological aspect of S4 is that the participant understands the contents of the exhibit after being approved by the presenter to join the conversation. When the participant or the bystander leaves the area, the current state returns to S1. The physical behaviors of state transitions are described in the following section.

2.3.2 Physical behaviors of state transitions

The behavior of transitioning from S1 to S2 is that a person looks at the presenter or the exhibit. We assume that

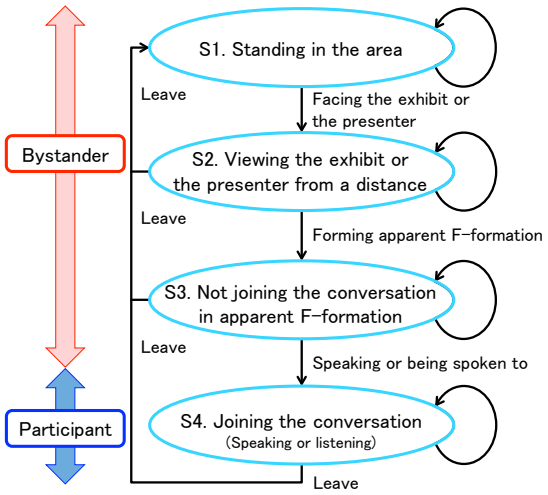

Figure 2. State transition model from bystander to participant in free-style conversational interaction, using physical behaviors.

facial direction is the same as gaze direction. To determine whether the behavior has occurred, we use the amount of change in facial direction. The behavior of transitioning from S2 to S3 is that a person moves nearer to the exhibit. We measure this as the change in distance between the person and the exhibit. The behavior of transitioning from S3 to S4 is that a person speaks to the presenter or is spoken to by the presenter. This behavior is measured as the amount of change in the voice direction. The behavior of transitioning from the current state (S2, S3, or S4) to S1 is that the person is not performing behaviors associated with a change to S2 or S3.

3. Our method for inferring the state transition

Here, we describe our method for automatically inferring the state transition using physical behaviors. We represent the center of the exhibit as $\boldsymbol{O}$, time is $t = 1, ..., T$, and the number of persons in the area is $i = 1, ..., I_a$ ($i = 1$ is the presenter, and $i \geq 2$ is the bystander or the participant). We represent the physical behaviors as a standing position vector $\boldsymbol{p}_i^t$, a facial direction unit vector $\boldsymbol{h}_i^t$, and an voice direction unit vector $\boldsymbol{v}^t$. The parameters of $\boldsymbol{p}_i^t, \boldsymbol{h}_i^t$ of the i-th person are acquired from cameras, and $\boldsymbol{v}^t$ is acquired from a microphone. We assign S1 to a person entering the area. The details of our method are described below.

3.1. Transition from S1 to S2

We determine whether the i-th person is viewing either the presenter or the exhibit. We first compute an angle between the i-th person and the presenter as

$$\theta_i^t = \cos^{-1}(\boldsymbol{p}_1^t - \boldsymbol{p}_i^t, \boldsymbol{h}_i^t)/||\boldsymbol{p}_1^t - \boldsymbol{p}_i^t|| \,, \tag{1}$$

where $\boldsymbol{p}_1^t$ is the standing position of the presenter, the origin of $\boldsymbol{h}_i^t$ is $\boldsymbol{p}_i^t$. We next compute the angle between the i-th person and the exhibit as

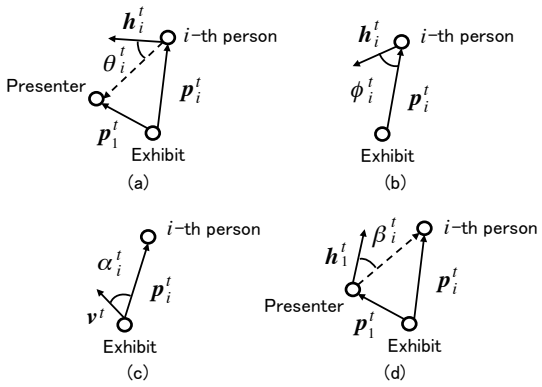

Figure 3. Parameters for inferring state transitions.

$$\phi_i^t = \cos^{-1}(-\boldsymbol{p}_i^t, \boldsymbol{h}_i^t)/||\boldsymbol{p}_i^t|| \,. \tag{2}$$

Figure 3 (a) and (b) shows the relationship between the parameters. Finally, we determine whether the i-th person is viewing either the presenter or the exhibit as follows:

$$\mathrm{check}_1(\theta_i^t, \phi_i^t) = \begin{cases} 1 & \min(|\theta_i^t|, |\phi_i^t|) < \eta_1 \,, \\ 0 & \text{otherwise} \,, \end{cases} \tag{3}$$

Where function $\min(\,)$ returns a minimum value, η_1 is a threshold. Our method returns a transition to S2 when check_1 returns 1; otherwise, the state remains as S1.

3.2. Transition from S2 to S3

In this transition, we determine whether the i-th person enters an apparent F-formation as follows:

$$\mathrm{check}_2(\boldsymbol{p}_i^t) = \begin{cases} 1 & ||\boldsymbol{p}_i^t|| < \eta_{2_1} I_f + \eta_{2_2} \,, \\ 0 & \text{otherwise} \,, \end{cases} \tag{4}$$

where I_f is the number of persons who enter in the apparent F-formation, and η_{2_1}, η_{2_2} are thresholds. The I_f is the summation of the number of bystanders of S3, the number of participants of S4, the i-th person, and the presenter. Existing methods [16, 12, 13] estimate F-formation using facial directions and body directions. We simply use standing positions and the number of persons because facial directions and body directions are virtually unchanged near the the exhibit in a poster presentation. Our method returns a transition to S3 when check_2 returns 1; otherwise, the state remains as S2.

3.3. Transition from S3 to S4

Here, we determine whether the i-th person speaks to the presenter or is spoken to by the presenter. Because the persons in apparent F-formation surround the exhibit, our

method uses voice direction $\boldsymbol{v}^t$ acquired from a microphone built into the exhibit $\boldsymbol{O}$. Note that our method does not allow for multiple persons speaking at the same time. We first consider the situation in which the i-th person speaks to the presenter. We compute the angle between the i-th person and the voice direction as

$$\alpha_i^t = \cos^{-1}(\boldsymbol{p}_i^t, \boldsymbol{v}^t)/||\boldsymbol{p}_i^t|| \ . \tag{5}$$

Figure 3 (c) shows the parameter α_i^t. We determine whether the i-th person speaks to the presenter as

$$\mathrm{check}_3(\alpha_i^t) = \begin{cases} 1 & |\alpha_i^t| < \eta_3 \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{6}$$

where η_3 is a threshold. We next consider the situation that the i-th person is spoken to by the presenter. We assume that the presenter looks at the face of the i-th person when speaking to him/her. We compute the angle between the i-th person and the facial direction as

$$\beta_i^t = \cos^{-1}(\boldsymbol{p}_i^t - \boldsymbol{p}_1^t, \boldsymbol{h}_1^t)/||\boldsymbol{p}_i^t - \boldsymbol{p}_1^t|| \ , \tag{7}$$

where $\boldsymbol{h}_1^t$ is the facial direction of the presenter. Figure 3 (d) shows the parameter β_i^t. We determine whether the i-th person is spoken to by the presenter as

$$\mathrm{check}_4(\alpha_1^t, \beta_i^t) = \begin{cases} 1 & |\alpha_1^t| < \eta_3 \wedge |\beta_i^t| < \eta_4 \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{8}$$

where α_1^t is the angle between the presenter and the voice direction, and η_4 is a threshold. Our method returns a transition to S4 when check_3 or check_4 returns 1; otherwise, the state remains as S3.

3.4. Transition from current state to S1

Here, we determine that the bystander or the participant is leaving the exhibit when he/she does not see the exhibit or the presenter, and moves some distance from the exhibit. Our method returns a transition to S1 when check_1 and check_2 return 0; otherwise, the state remains at its current value.

4. Experiments

4.1. Evaluating thresholds of our method

4.1.1 Setup for acquiring video sequences

We acquired video sequences of bystanders and participants. We simulated a poster presentation and used the six devices (Microsoft Kinect v2) K_1 to K_6, as illustrated in Figure 4 (a). The exhibit was a 26-inch display with a exhibit title. The presenter explained the presentation slides to the participants, using the display. We changed the slide

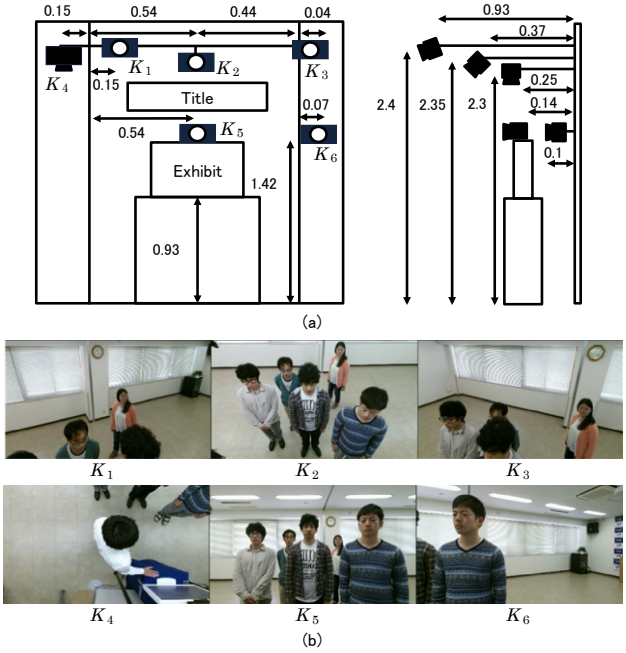

Figure 4. (a) Setup for acquiring video sequences [m]. (b) Examples of images acquired at the same time.

each time. We set up cameras at 30 fps and the microphone at 16k Hz embedded in a Kinect device. Figure 4 (b) shows examples of acquired images from the devices. We used two setups, with and without partitions to separate booths. The size of a booth was 2.1×2.1 m^2. Note that we used a body-attached sensor (MicroStrain 3DM-GX3-25) for the presenter only, to acquire facial direction.

We computed a rotation matrix and a translation vector of each Kinect device, using a calibration method [17]. The origin of the world axis coordinate was the center of the exhibit $\boldsymbol{O}$. To determine $\boldsymbol{p}_i^t$, we integrated head positions acquired from Kinect devices by taking the average of the positions that were within 0.1 m at time t. To determine $\boldsymbol{h}_i^t$, we took the average of the directions acquired from the Kinect devices for $\boldsymbol{p}_i^t$. To determine $\boldsymbol{v}^t$, we used the microphone in K_5.

We collected 15 subjects (average age 21.7 ± 0.9) and randomly nominated a presenter, a bystander, and participants. We gave the bystanders the following tasks:
T1: Exiting the area immediately after noticing the exhibit,
T2: Exiting the area after viewing the exhibit from a distance,
T3: Exiting the area after moving closer to the exhibit,
T4: Exiting the area after waiting to speak to the presenter but not joining to the conversation,
T5: Exiting the area after speaking to the presenter,
T6: Exiting the area after being spoken to by the presenter.
We set I_a as 3 or 6 persons randomly selected from the 15

subjects. In each task, we acquired four video sequences by changing the subjects selected. We collected 96 video sequences (total 70 minutes) to evaluate thresholds of our state transition model by covering the all paths.

4.1.2 Evaluation results of the thresholds

To evaluate the performance of inferring state transitions, we used the F-measure that is the harmonic mean of precision and recall. By comparing the manually labeled state and the inferred state on each time, we counted true positives, true negatives, false positives, and false negatives for each state. We applied the judgments of state, using check_1 to check_4 30 times per second. To address the difference of time length between video sequences, we randomly sampled the judgments when computing F-measure.

Two annotators labeled states at all times in all video sequences by discussing between themselves. When their opinions of the labeled states differed, they interviewed the subjects to determine which state was appropriate. We also checked the state labels of different annotators who were not completely familiar with our state transition model. We observed that their state labels were in near-uniform agreement.

We evaluated the thresholds of our method. Figure 5 shows the average and the standard deviation of the F-measure when a certain threshold was fixed and other thresholds were changed. We used Bonferroni's method for multiple tests ($p < 0.01 : **$, $p < 0.05 : *$). High inference performance occurred when η_1 was 15 degrees or more, η_{2_2} was between 1.55 meters and 1.85 meters, η_3 was 15 degrees or less, and η_4 was 9 degrees or less. We believe that controlling η_{2_1} by using the number of persons in apparent F-formation was effective because performance was higher when η_{2_1} was larger than 0. The maximum F-measure was 0.749, using $\eta_1 = 15, \eta_{2_1} = 0.1, \eta_{2_2} = 1.55, \eta_3 = 15, \eta_4 = 9$.

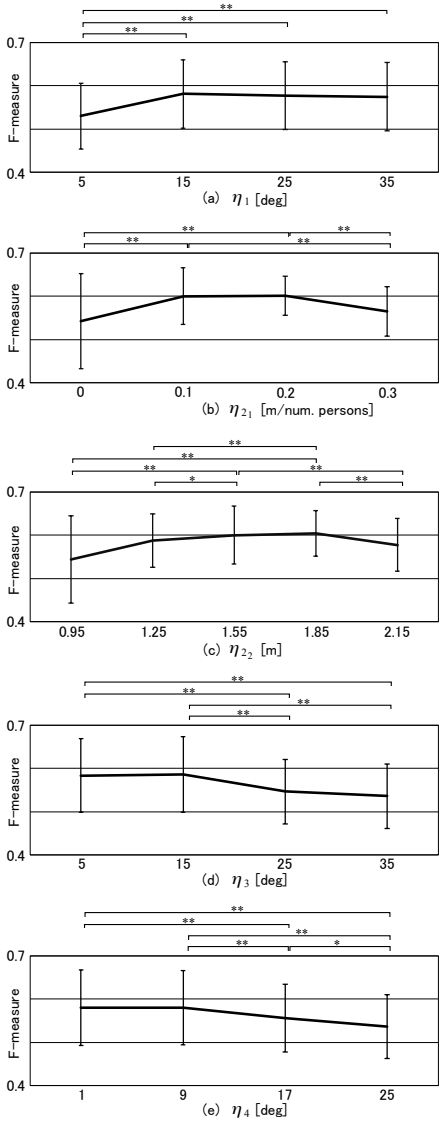

Figure 5. Average F-measures evaluated on C1 to C6 when a certain threshold is fixed and the other thresholds are changed.

4.2. Evaluation of state transition inference in free-style conversational interaction

We evaluated our method using a video sequence of free-style conversational interaction. We used three exhibits and presenters, as illustrated in Figure 6 (a). We collected data from another six subjects who were not completely familiar with our state transition model. We asked these six subjects to visit the three exhibits in 600 seconds. Figure 6 (b) shows examples of images acquired from the Kinect devices set in Exhibit 2. We used the top two thresholds with the highest inference performance in Section 4.1.2.

Figure 7 shows the average and the standard deviation of the F-measure when a certain threshold was fixed and the other thresholds were changed. We applied Welch's t-test for η_{2_1}, η_{2_2}, and the Student's t-test for η_1, η_3, η_4. Significant differences emerged ($p < 0.01$) on η_{2_1}, η_{2_2}, and η_4. The maximum F-measure was 0.654 when $\eta_1 = 25$, $\eta_{2_1} = 0.1$, $\eta_{2_2} = 1.55$, $\eta_3 = 15$, $\eta_4 = 9$. We also obtained almost the same performance (0.653) using $\eta_1 = 15$. We believe that our method works effectively in free-style conversational interaction.

5. Conclusion

We proposed a method for automatically inferring the state transition from bystander to participant in free-style conversational interaction. We designed our model using physical behaviors in the context of a video sequence of observations of a poster presentation. Experimental results show that the use of parameters of physical behaviors and

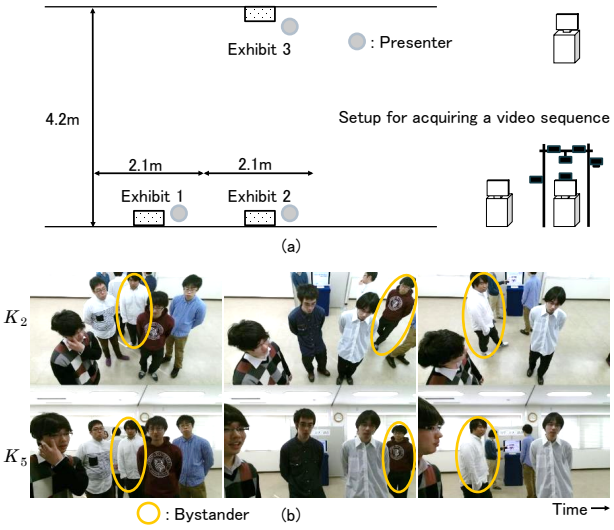

Figure 6. (a) Setup for acquiring a video sequence of free-style conversational interaction [m]. (b) Examples of camera images of free-style conversational interaction.

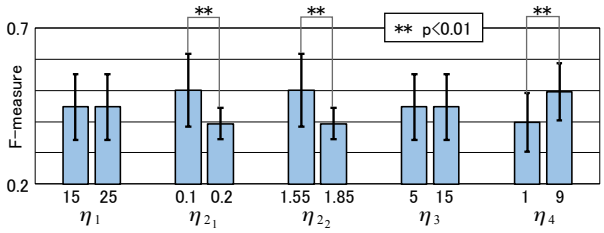

Figure 7. Performance (F-measure) in inferring states while bystanders and participants are freely moving.

our method worked well. In future work, we intend to develop a method for considering the relationship between bystanders and participants.

Acknowledgment This work was partially supported by JSPS KAKENHI Grant No. JP25220004.

References

[1] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe. Analyzing free-standing conversational groups: A multimodal approach. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 5–14, 2015.

[2] M. Bono, N. Suzuki, and Y. Katagiri. An analysis of participation structures in multi-party conversations. *Cognitive Studies*, 11(3):214–227, 2004.

[3] L. Chen, R. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. Han, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, R. Tuttle, and T. Huang. Vace multimodal meeting corpus. In *Machine Learning for Multimodal Interaction*, pages 40–51, 2006.

[4] H. Clark. *Using Language*. Cambridge University Press, 1996.

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

[6] H. Hung and B. Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 231–238, 2011.

[7] K. Inoue, Y. Wakabayashi, H. Yoshimoto, and T. Kawahara. Speaker diarization using eye-gaze information in multi-party conversations. In *Proceedings of Interspeech*, pages 562–566, 2014.

[8] A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, 1990.

[9] D. McNeill. Gesture, gaze, and ground. In *Machine Learning for Multimodal Interaction*, pages 1–14, 2006.

[10] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41(3):409–429, 2007.

[11] M. Poel, R. Poppe, and A. Nijholt. Meeting behavior detection in smart environments: Nonverbal cues that help to obtain natural interaction. In *Proceedings of 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6, 2008.

[12] E. Ricci, J. Varadarajan, R. Subramanian, S. R. Bulo, N. Ahuja, and O. Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *Proceedings of International Conference on Computer Vision*, pages 4660–4668, 2015.

[13] R. Subramanian, J. Varadarajan, E. Ricci, O. Lanz, and S. Winkler. Jointly estimating interactions and head, body pose of interactors from distant social scenes. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 835–838, 2015.

[14] Y. Sumi, M. Yano, and T. Nishida. Analysis environment of conversational structure with nonverbal multimodal data. In *Proceedings of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 44:1–44:4, 2010.

[15] T. Tung, R. Gomez, T. Kawahara, and T. Matsuyama. Multiparty interaction understanding using smart multimodal digital signage. *IEEE Transactions on Human-Machine Systems*, 44(5):625–637, 2014.

[16] S. Vascon, E. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, 143:11–24, 2016.

[17] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.