Automatic Gaze Analysis in Multiparty Conversations based on Collective First-Person Vision

Shiro Kumano, Kazuhiro Otsuka, Ryo Ishii, and Junji Yamato

***Abstract*— This paper extends the affective computing research field by introducing first-person vision to automatic conversation analysis. We target medium-sized-party face-to-face conversations where each person wears inward-looking and outward-looking cameras. We demonstrate that the fundamental techniques required for group gaze analysis, i.e. speaker detection, face tracking, and gaze estimation, can be accurately and effectively performed via self-training in a unified framework by gathering captured audio-visual signals to a centralized system and using a general conversation rule, i.e. listeners look mainly at the speaker. We visualize the characteristics of participants' gaze behavior as a gazee-centered heat map, which quantitatively reveals what parts of the gazee's body and for how long the participant looked at it while the gazer speaks or listens. An experiment involving two groups of six-person conversations demonstrates the potential of the proposed framework.**

I. INTRODUCTION

Face-to-face conversation is the primary way of sharing information, understanding others' emotion, and making decisions in social life. Accordingly, to develop conversational agents or computer-mediated telecommunication, automatic meeting analysis has been acknowledged as a basic research area [1], [2]. Most previous studies offer preliminary steps toward the recognition of the verbal/nonverbal behaviors of conversation participants, including speech, gaze, facial expressions, gestures and postures. They use data captured by microphones and cameras stationed in the environment.

The emotional aspect in face-to-face conversations is now being addressed mainly in dialogues [3], [4], [5], [6]. To develop a model that can automatically infer emotion from behavior, human annotations are often required as the ground truth; participants' self report or external observer's judgment regarding emotion, e.g. valence-arousal dimensions, and coding of nonverbal behavior, including gaze, facial expression, and head gesture. The automatic assessment of nonverbal behaviors with high accuracy is still challenging for multiparty conversations, in particular for medium-(five- to ten-participants) or large conversations. In these scenarios, the face directions against the fixed cameras varies largely when people to look at other participants. Thus, it is hard for state-of-the-art face analyzers, e.g. CERT [7], or even non-expert human coders, to fully distinguish subtle gaze/facial actions and head movements.

S. Kumano, K. Otsuka, R. Ishii and J. Yamato are with NTT Communication Science Laboratories, 3-1 Morinosato-Wakamiya, Atsugi, Kanagawa, Japan. kumano@ieee.org

© 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. Find the published version of this article under http://ieeexplore.ieee.org/Xplore/home.jsp.

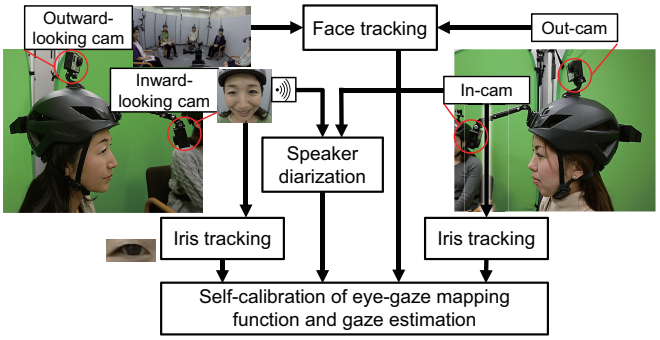

Fig. 1. Flow of collective FPV for self-calibrating gaze analysis

Our solution is to introduce first-person vision (FPV), where the wearer's field of view (FOV) is captured by an outward-facing camera (out-cam in short), together with his/her face by an inward-facing camera (in-cam in short). FPV is inherently superior for measuring some conversational behaviors of the wearer because the facial motions and head motions are completely separated by these cameras. The wearer's face is nearly always stable in the in-cam image, as in [8], while any head motion yields large motion flow in the out-cam image; this is useful for head gesture recognition.

Among these nonverbal behaviors, one of the most fundamental ones is gaze given its importance in several social functions [9], e.g. monitoring, visual feedback, expressing emotion/empathy [10], and regulating the flow of the conversation. However, even when using glass-type eye trackers, the collection of accurate group gaze behavior, e.g. who is looking at whom, which facial/body part, and when, is still time-consuming due to the requirement of manual intervention for calibrating the gaze trackers, localizing faces in the FOV images, and identifying the speaker at every moment. Most previous conversation analyses actually targeted dialogues [11], trialogues [12], and four-party [13] conversations, or used head poses in medium-sized conversations as rough estimates of visual focus-of-attention, e.g. [14].

Our basic idea is to combine the fundamental CV techniques required for these analytical steps to develop a (nearly-)fully automatic system that offers deeper analysis. Targeting a likely new conversational scenario, where each interlocutor has his/her own in- and out-cams with micro-

phones attached to a rigid worn object, e.g. a helmet or glasses, we apply two established frameworks: The first is inter-sensor collaboration, where captured signals are gathered to a centralized system, or shared among distributed collaborative systems [15]. The second is the active use of the primary characteristic of multi-party conversation elucidated so far; listeners in conversation look mainly at the speaker [9], [12], [16]. We call this framework, which subsumes these two frameworks, *Collective First-Person Vision (Co-FPV)*.

The contributions of this paper are as follows: 1) Co-FPV is proposed for estimating the gaze behavior of each interlocutor in a multi-party conversation via self-calibration. 2) *A conversational rule* is introduced for the self-calibration of an eye-gaze mapping function that transforms the iris position in the in-cam to the gaze point in the out-cam. 3) Face tracking via inter-sensor collaboration in the Co-FPV framework is proposed. These techniques realize the automatic characterization of the gaze behavior of participants as a *gazee-centered* heat map, which reveals which part of the gazee, and for how long, the participant looked at it. We demonstrate that Co-FPV offers promising performance despite its simplicity. Fig. 1 illustrates the proposed framework.

The remainder of this paper is organized as follows. Section II describes related work. Section III reports a preliminary gaze behavior experiment. Section IV explains the proposed framework. Sections V and VI details the experiment, the results, and a discussion. Finally, our summary and future work are provided in Section VII.

II. Related Work

This section positions this study by comparison with related work as regards the following four topics: gaze behavior in conversation, first-person vision, gaze analysis with eye tracker, and eye tracker self-calibration.

A. Gaze behavior in conversation

Gaze offers several conversational functions, as described in Section I. Of particular note for this study is that people pay attention by orienting their gaze toward the speaker [11]. These tendencies are also revealed in three-party [12] and four-party [13] conversations. We demonstrate a similar trend for conversations involving more participants, six-party conversations, in III.

B. First-person vision (egocentric vision)

FPV is a hot topic in the computer vision community, and the number of related papers is rapidly increasing [17]. Several tasks have already been tackled: e.g. gaze tracking, activity recognition, three-dimensional reconstruction, and video summarization [18]. Of particular note is the pioneering work that targets social interactions in FPV. In [19], the type of social interaction, i.e. monologue, dialogue or discussion, is classified from the poses of people in the images captured by a camera being worn by a person. A similar task is tackled in [20] by combining egocentric images with images by a stationed camera in the environment. However, these studies do not consider situations where everyone has his/her own wearable camera(s).

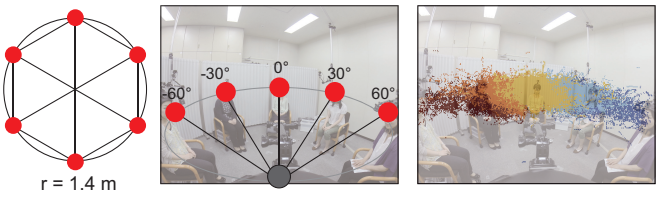

Fig. 2. Interlocutor arrangement, left) top view and middle) egocentric view, and right) spatial distribution of face position in the out-cam images. Colors indicate different locations relative to the wearer.

C. Gaze analysis with eye tracker

The use of eye trackers to understand how human beings observe photographs or movies has long been a subject of research. A good example involves investigating the difference in gaze behavior between typically developed people and people with autism, e.g. [21]. Another topic involves building a computational/stochastic model that well explains human gaze patterns. A variety of models have already been proposed, e.g. Itti and Koch's model [22] of low-level saliency, and a high-level objective model. For example, it is reported in [16] that while observing a movie depicting social interaction, people tend to look at the turn holder, and low-level saliency models fail to explain such gaze patterns. The main drawback of these studies is that they often rely on human intervention for eye-gaze calibration, and face and speaker detection. The proposed framework automates these processes all together. This makes it easier to analyze the gaze patterns of both observers of a social interaction and *people involved in that social interaction*.

D. Self calibration of eye-gaze mapping function

Various self calibration techniques have already been proposed [23]. Some recent studies on passive vision-based eye trackers are based on the prediction of the gaze point by using a low-level visual saliency model, e.g. [24], or use others' gaze patterns to target images [25]. However, the former fails when viewing social interaction, and the latter is inapplicable to unknown conversational scenes. The current study differs from those studies mainly in the sense that conversational saliency, namely turn-taking, is used as prior knowledge with which to predict the gaze point of a target person without using the gaze behavior of others.

III. Preliminary Gaze Behavior Experiment

This paper targets medium-sized, six-party, conversations, unlike previous studies explained in II-A. Thus, we first report a preliminary experiment and its results to grasp the notable characteristics of gaze behavior for self-calibrating gaze analysis. Moreover, we assume that people are sitting in a circle, as shown in the left part of Fig. 2, and at most one speaker exists at any given time. In the preliminary experiment, the gaze of each person and the speaker were annotated by one person. The details are provided later in Section V-B.

Fig. 3. Prototype camera mount. Left: In- and out-cams (GoPro Hero3+ x 2) are attached to a lightweight mountain climbing helmet. Right: Captured images of those cameras.

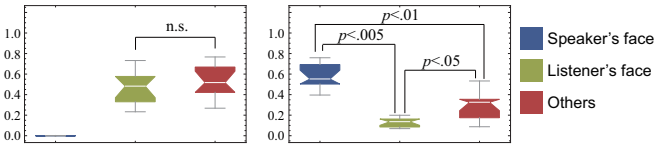

Fig. 4. Gaze targets of speaker (left) and listener (right). The most salient gazee is the speaker for listener.

A. Measurement device

Fig. 3 shows our prototypical measurement device in this study. Note that designing a smart hardware is out of the main scope of this paper. The total number of cameras is $2N$, where N is the number of people in the social interaction. $N = 6$ in this study. The spatial resolutions of the cameras, i.e. the number of pixels and FOV angles, are assumed to be already known.

B. Validity of our basic assumption

Fig. 4 shows the frequency at which speakers and listeners were looked at. As in previous studies [9], [12], the listeners more frequently looked at the speaker than at another listener (paired t-test with Bonferroni correction, $t(11) = 12.8$, $p < .005$, $r = .99$), or other target ($t(11) = 4.0$, $p < .01$, $r = .94$). These tests used the probabilities that each participant was looking at them. In comparison, the speakers looked at the listeners to the same extent as other targets ($t(11) = .61$, $p > .05$, $r = .39$); at worse, it was not clear who the gazee was among the listeners. These results suggest that eye-gaze self-calibration can be realized by identifying the iris center position of the wearer in the in-cam image, the speaker that the wearer was listening to, and the face position of the speakers in the out-cam image.

C. Spatial characteristics of our data

Right part of Fig. 2 shows the spatial distribution of others' faces in the out-cam images. Although specific in our settings and unusual in other settings, e.g. standing or in-motion conversations, the face positions cover the full range of the horizontal axis, but only a narrow range of the vertical axis. These characteristics originate from the fact that the people were sitting on adjacent chairs. Although the wearer's head sometimes moves vertically when nodding or looking upward or downward, it is difficult to cover the entire vertical range with this setting.

The spatial distribution of the faces suggests that even if we know precisely the gazee that the wearer is looking at every moment, the number of samples of the eye-gaze calibration will be biased, i.e. large on the horizontal axis, but small on the vertical axis. To enhance the accuracy along the y-axis, joint calibration for both directions is advantageous. Moreover, the horizontally spread distribution in the mid-level of the image suggests that the radial distortion of the out-cam is insignificant, because it mainly affects the upper and lower parts of the image.

IV. Proposed Self-Calibrating Gaze Analysis

This section describes the proposed self-calibrating gaze analysis for Co-FPV. Fig. 1 summarizes the flow. It mainly consists of three steps. Step A) First, faces in the out-cam images are tracked by the proposed two-stage coarse-to-fine Co-FPV analysis. Step B) The eye-gaze mapping function is then self-trained by assuming that the listener is looking at the speaker's face. The speaker is identified by using the power of the audio signals obtained from everyone's in-cam. Step C) Next, the gaze point of each person in the out-cam images is estimated by using the iris center and the trained eye-gaze mapping function. Finally, a gazee-centered heat map of each wearer is developed through gazee recognition.

A. Fine face localization in out-cam images

This subsection explains how we obtain, in an out-cam image of the target wearer $p \in \{1, \cdots, N\}$, the face center of each person, $q \in \{1, \cdots, N | q \neq p\}$, $\boldsymbol{m}_q^{(p)}$. The face center is defined as the mid-point between the eyes. Hereafter superscript p is omitted because we consider that the target wearer is p. We assume that a rough estimate of the face location of each person in the image is already known by using an object detection/tracking technique, described in Section V-A, in the first stage. The aim of the second step is to refine the face center position.

1) Overview: We obtain the face center in out-cam images precisely by rotating a three-dimensional frontal face template. In previous FPV studies, the head rotation angle is often estimated from the out-cam image of *the target wearer*, p, e.g. [19]. On the other hand, we obtain person q's head pose from q's *own* out-cam image. Fig. 5 shows the flow of the proposed face localization.

Our approach is based on the fact that q's head pose to p has a one-to-one correspondence (nearly linear) with p's position in q's out-cam image. For example, if q's face and the out-cam are facing in the same direction and p is located in the middle (or left) part of q's image, then it means that q is directly facing p (or is facing the right side of p), as shown in Fig. 6. This approach is much more useful than the traditional approach. Consider the case where the front of q's face, width of 30 pixels in p's image, is rotated horizontally by one degree. If the camera horizontal FOV is 1920 pixels and 122.6 degrees, the face center shifts by 16 (=1920/122.6) pixels in q's image (the proposed approach), while the center

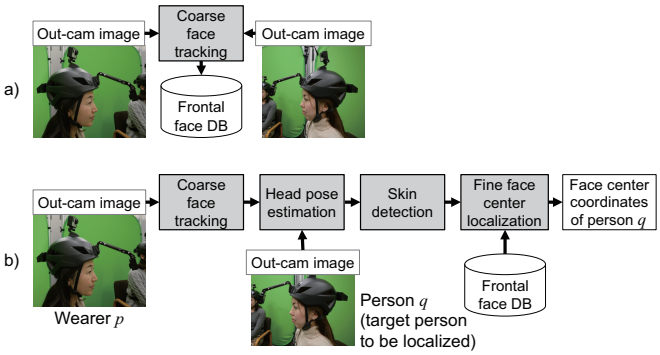

Fig. 5. Flow of the proposed face center localization: a) First, frontal face images with eye positions are collected continuously from the images of everyone's out-cams. b) Then, the face center coordinates of person q in the out-cam image of wearer p are determined by using person q's out-cam image and the frontal face database.

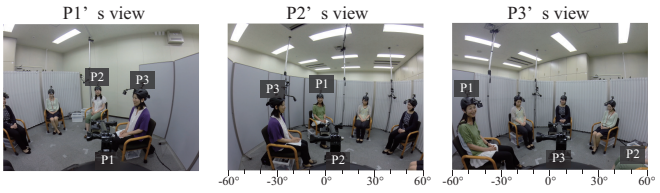

Fig. 6. Relation between facial pose and position in out-cam images. P1's face positions in the images of P2 (middle) and P3 (right) provide the face poses of P2 and P3 in P1's image (left). In this case, the horizontal angles of P2 and P3 relative to P1 are close to -5 and -50 degrees, respectively.

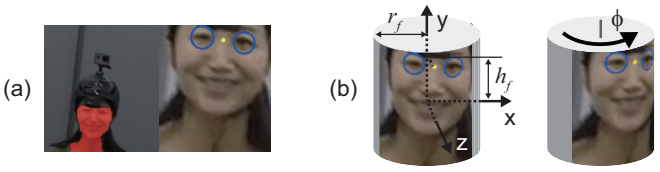

Fig. 7. a) Detected skin regions (red masks), the centroid of both eyes, i.e. face center, (yellow dot) in frontal faces. b) The face center is mapped onto a cylinder, and rotated according to estimated head pose.

moves by only 0.3 (=$30/2 \cdot \sin(1°)$) pixels in p's image (the traditional approach).

2) Facial model: The face model of each person consists of a three-dimensional shape and the position of the face center on the shape. Face models are automatically generated selectively from images captured by the out-cams of all the people that satisfy the following conditions: First, the captured faces are almost frontal. The images are not necessarily captured by p's out-cam. Second, both eyes are detected in the image by an eye detector. The position of the face center is obtained as the mean position of the centroid of the eyes in the selected images.

As a face shape model, we use a cylinder, like [26], with a radius of r_f. The face center is assumed to be positioned on its surface. The directions of the face coordinate system of the axes are horizontal, vertical and facial-frontal, as shown in Fig. 7. The face center coordinate is $\boldsymbol{x}_c = (0, h_f, r_f)^{\mathrm{T}}$. Height h_f is the mean y-coordinate of the mid-point between the two eyes in the bounding boxes of skin region. Radius r_f is set at the mean of the half width of the boxes, w. The skin region is detected by color-based skin masking in the HSV space around the rough estimate of face position obtained in the first step, the red areas in Fig. 7.

3) Face localization in out-cam images: A refined face center coordinate in the out-cam is obtained as

$$\hat{\boldsymbol{m}} = f_p(\boldsymbol{R}\boldsymbol{x}_c) + \bar{\boldsymbol{m}}, \tag{1}$$

where $\boldsymbol{R}$ denotes a three-dimensional rotation matrix, and $\bar{\boldsymbol{m}}$ is the image coordinate of the center of the skin region. As camera model f_p, we use a weak-perspective camera, i.e. $f_p(\boldsymbol{x}) = (x, y)^{\mathrm{T}}$, where $\boldsymbol{x} = (x, y, z)^{\mathrm{T}}$.

B. Self-training of eye-gaze mapping function

The objective of eye-gaze calibration is to find a mapping function, f, that associates iris center coordinates in the in-cam, $\boldsymbol{e} = (e_x, e_y)$, with the corresponding gaze point in the out-cam, $\boldsymbol{g}$, namely $\boldsymbol{g} = f(\boldsymbol{e}; \Theta)$, where Θ denotes the parameters of f. The proposed self-calibration assumes that the gaze point is the face center of speaker s at that moment, i.e. $\boldsymbol{g} = \hat{\boldsymbol{m}}_s$, when the wearer p is not the speaker.

Parameters Θ are obtained by minimizing the following objective function:

$$\Theta = \arg\min_{\Theta} \Sigma_j dist(f(\boldsymbol{e}_j; \Theta), \boldsymbol{g}_j) + \lambda(\Theta - \bar{\Theta})^2, \tag{2}$$

where $dist$ is a distance function, and j denotes a sample index. The second term is a regularization term that restricts Θ so that it remains around $\bar{\Theta}$, a rough estimate of Θ.

1) Mapping function f: To evaluate the basic validity of our framework, this paper performs model-based two-dimensional eye-gaze calibration. Note that the form of the mapping function is not the main focus of this study, and different forms of mapping functions to suit different hardware designs are applicable to the proposed framework.

First, we impose several assumptions to simplify the mapping function. By ignoring gaze parallax, i.e. assuming that both eyes are always oriented in the same direction, we simply consider the centroid of the eyes. Second, the following geometrical parameters are considered: eyeball radius r, the distance between the center of the eyeball and the focal point of the in-cam, d, and the yaw and pitch angles of the in-cam relative to the eyeball, θ_x and θ_y, respectively. The remaining geometrical parameters, i.e. the locations of the centroid of the eyeballs and the focal point of the out-cam are the same, the roll angles of both the in- and out-cams, and the yaw and pitch angle of the out-cam relative to the eyeball are assumed to be zero by the following pre-processing: The in-cam is aligned as both eyes lie on a horizontal line and the eye centroid is located at the image center. The out-cam is aligned as the faces of the person in front of the wearer is located at the image center. Fig. 8 shows the geometrical relationship between the cameras and the eyeball.

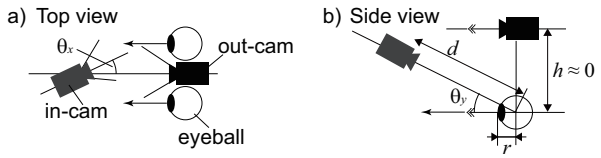

Fig. 8. Assumed geometrical relationship of cameras and eye

In this model, mapping function f is approximated by[1]

$$f(\boldsymbol{e}) = \alpha_{\mathrm{o}}[\arcsin\{d/r \cdot \tan((-e_x, e_y)^{\mathrm{T}}/\alpha_{\mathrm{i}})\} + (\theta_x, \theta_y)^{\mathrm{T}}], \quad (3)$$

where α_i and α_o are the scale factors of the in- and out-cams that relate degrees to pixels in both axes, respectively, namely the ratio between the degrees and the number of pixels of the camera's FOV. The sign of e_x is changed because the x-axis is flipped in the in-cam, as shown in Fig. 8.

Assuming that $\boldsymbol{e}$ is close to $(0, 0)$ in (3), we can approximate mapping function f as a similarity transformation:

$$\boldsymbol{g} = f(\boldsymbol{e}) \approx \bar{f}(\boldsymbol{e}) = \begin{pmatrix} -a & 0 & b \\ 0 & a & c \end{pmatrix} \boldsymbol{e}', \quad (4)$$

where $\boldsymbol{e}'$ denotes the augmented vector of $\boldsymbol{e}$, and $\Theta = (a, b, c)$. This makes the training more robust. This approximation introduces large errors when the gaze direction is roughly over 60 degrees, which is infrequent in our settings, or when the camera is very close to the face and has significant radial distortion, e.g. for glass-type devices. To avoid such errors, the exact form in Footnote 1 is superior.

2) Sample selection: Training samples are selected that satisfy the following three conditions: the speaker's face, fixation, and inlier. Fixations are detected based on the dispersion as a short temporal block with small variation in $\boldsymbol{e}$ without blinking; this is the dispersion-threshold identification [27]. Inliers are determined by $|\bar{f}'(\boldsymbol{e}) - \hat{\boldsymbol{m}}_s| < \tau_e$, where $\bar{f}'$ is identical to $\bar{f}$ except for $a = \bar{a}$ and $b = c = 0$ in $\bar{f}'$. Cases where the wearer is looking at another listener can be removed by this thresholding. Outlier removal is based on the geometry-based calibration, which gives a prior knowledge of parameter $\bar{a}$.

These constraints are not perfect and sometimes yield incorrect samples. Accordingly, we eliminate their effect by using, as a distance function in (2), a robust function $dist(\boldsymbol{m}_1, \boldsymbol{m}_2) = |\boldsymbol{m}_1 - \boldsymbol{m}_2|^2/(\kappa^2 + |\boldsymbol{m}_1 - \boldsymbol{m}_2|^2)$. Note that other robust estimation techniques, e.g. the Random Sample Consensus (RANSAC), are also applicable.

C. Estimation of gaze point and gazee

After mapping function f is trained, the gaze point at each time, $\hat{\boldsymbol{g}}$, is obtained by substituting the iris center coordinates $\boldsymbol{e}$ at that time into (4). The gazee, $\hat{q}$, is identified as the person nearest to the gaze point as determined by Euclidian distance. Moreover, if the distance exceeds threshold τ_d, the wearer is not considered to be looking at anyone's face.

We generate the gazee-centered heat map as a (relative) gaze duration heat map [28], which shows the accumulated time the wearer spent looking at the different areas of the other interlocutors. The estimated gaze point $\hat{\boldsymbol{g}}$ in the gazer's coordinates is mapped into gazee $\hat{q}$'s face coordinate system as: $\boldsymbol{\xi} = 2/w \cdot (\hat{\boldsymbol{g}} - \hat{\boldsymbol{m}}_{\hat{q}})$, where w is the width of the bounding box of the detected face, described in Section IV-A. Heat maps are then obtained by collecting $\boldsymbol{\xi}$ values during gaze fixation and visualizing the density of the collected $\boldsymbol{\xi}$ values.

V. Experiment

This section describes the experiment conducted to evaluate the performance of the proposed method.

A. Fundamental techniques

The TLD tracker [29] was used to obtain the coarse location of faces in the out-cam images. It was initialized by manually assigning each person's face region, which included the neck and the bottom part of the helmet. The tracker roughly but quite robustly detected the faces, even though the initialized faces were often non-frontal and blurred, and the left- and right-most persons in the image repeatedly appeared/disappeared from the FOV during conversation because of the wearer's head rotation in our settings. Initialization is the only manual intervention needed when employing the proposed method; this can be automated by the multi-view face detector, introduced in [30].

The in-cam images were aligned in advance as the horizontal coordinates of both eyes are the same and their centroid is located at the center of the image. This removes occasional slight helmets shift during conversation. The images were aligned in the following steps: First, both eyes were localized by an eye detector [31]. Then, the eye positions were smoothed with a temporal filter to compensate errors. Finally, the translation vector and rotation matrix were calculated and applied to the images.

To localize iris centers in the eye-aligned images, we used Invariant Isocentric Patterns (IIPs) [32]. Although the original paper [32] found the iris centers of both eyes independently, this paper jointly localizes the iris centers by imposing the constraint that their horizontal coordinates should be the same and their separation is constant. This hampers gaze estimation in the depth direction, but makes it robust in the horizontal and vertical directions. Blinking was determined by thresholding the vertical coordinate of the upper eyelashes which were detected as a dark region.

Speakers were identified as the person generating the maximum acoustic power in everybody's in-cams at each moment. If the maximum power is lower than threshold τ_a, no one is assumed to be speaking. This ignores overlapping speech among interlocutors; in practice, simultaneous utterances occasionally occurred in the discussion sessions. However, the results show that their impact is limited.

[1] This equation is obtained as follows. Considering x-axis in Fig. 8, gaze angle ψ changes the iris center position at $r\sin(\psi - \theta_x)$ in the physical space, while the change is $(d - r\cos\theta_x)\tan(-e_x/\alpha_i)$ pixels in the in-cam image in weak-perspective projection. Assuming $d \gg r$ approximates ψ as $\arcsin(d/r \cdot \tan(-e_x/\alpha_i))$. The gaze shift in the out-cam image is $g_x = \alpha_o\psi$. Linking these equations with regard to ψ yields $g_x = \alpha_o[\arcsin\{d/r \cdot \tan(-e_x/\alpha_i)\} + \theta_x]$, i.e. (3). The derivation for y-axis is equivalent except for the signs of e_x and e_y.

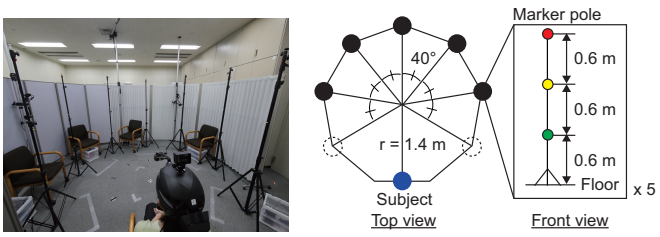

Fig. 9. Marker-based calibration scene for performance evaluation

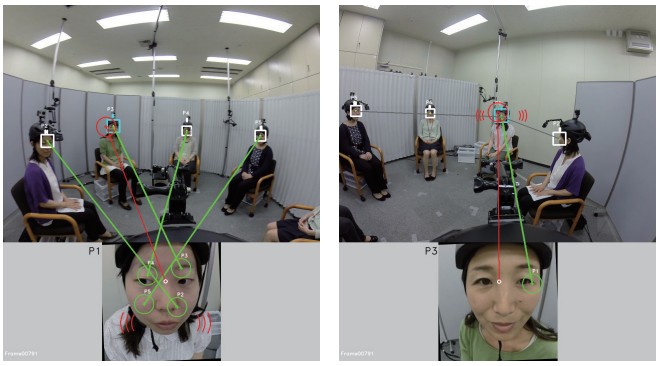

Fig. 10. Typical estimation results: left) speaker's view and right) the gazee's view. The in-cam images of the wearers are superimposed at the bottom after horizontal flipping. Red circles indicate the gaze points of the wearers. Green and gray circles/lines denote the others' gaze to the wearer or another, respectively. Cyan and white boxes are tracked faces; cyan indicates the gazee of the wearer. Red waves denote the speaker. This movie is available from http://www.brl.ntt.co.jp/people/kumano/emospace2015/.

Alternatively, it is possible to employ person-wise binary thresholding to decide whether each person is speaking or silent.

B. Conversation settings

Twelve Japanese women in their twenties to forties participated in this experiment. They were divided into two six-person groups. Members of the same group had not met before the experiment. Both groups were instructed to engage in two conversations and did so. First, each member introduced themselves taking about 1.5 min for each introduction. The groups then held discussions and built a consensus as a group, i.e. they agreed on a single answer, related to a given topic within 10 min. The radius of their circular arrangement was 1.4 m, as shown in Fig. 2.

Noticeable conversational characteristics are: 1) in the self-introduction session, the subjects spoke simply in turn. Consequently, there was a single speaker at any given moment, and the listeners mainly looked at the speaker, as we assumed and as reported in some papers [9], [12]. 2) The discussion session was more challenging for our task; this conversational rule was often violated. Overlapping speech often occurred as a result of listeners' backchannel behavior, and listeners occasionally looked at other listeners to understand the conversation situation for obtaining a turn.

C. System settings

The spatial resolution of the out-cams was set at 1920×1440 pixels (122.6×94.4 degrees), while that of the in-cams was set at 848×480 pixels (118.2×69.5 degrees). The temporal resolution of all cameras was set at 30 fps. The $2N$ cameras were synchronized by starting them simultaneously by remote control. Mapping function f was trained separately for each person and each target conversation session to avoid the severe drift caused by helmet shift. Only slope a was regularized in (2). Its rough estimate $\bar{a}$ is obtained as $\alpha_{\rm o}\bar{d}/\alpha_{\rm i}r$ from the first-order derivative of (3) at $\boldsymbol{e} = (0, 0)^{\rm T}$; $\bar{d}$ is a rough estimate of d. In (1), the head was assumed to rotate only along the vertical (y) axis by angle ϕ, because horizontal head rotations were frequent and large while vertical and in-plane rotations were infrequent and relatively small.

The model parameters were set as follows: Eyeball radius r was set at 1.25 cm with reference to [33], and $\bar{d} = 17$ cm (this yields $\bar{a} = 30$). Threshold τ_e was set at 200 pixels, and $\kappa = 100$ and $\lambda = 0.01$. Threshold τ_d was set at 80 pixels (= 5 degrees) with reference to [34], which investigated the range of the human gaze point while looking at another person. The original audio signals were normalized to the zero-mean-unit-variance in a pre-processing step, and τ_a was set at half of the mean power for each person.

D. Manually-generated data for performance evaluation

We manually prepared three types of validation data. The first were samples for assessing the estimated gaze points. The subjects were, after the conversation sessions, asked to look at a specified physical marker placed in space, as shown in Fig. 9. Five (horizontal) $\times$ three (vertical) markers were used in this study[2]. The marker positions in the out-cam images were localized by an annotator. Second, to assess the error yielded by the fundamental techniques, the annotator assigned the image coordinates of the face center in the out-cams and of the iris center in the in-cams. 374 face- and 360 eye-images were randomly selected. Third, to evaluate the performance of gazee and speaker recognition, the annotator also gave them frame-by-frame labels throughout the four conversation sessions.

E. Accuracy assessment

Fig. 10 and Table I and Table II show the performance of the proposed framework, which generally worked well for each task. Gaze point estimation errors were obtained by using the physical markers, while the remaining performances were obtained by using only the conversation data.

1) Face center localization: The proposed framework greatly improves the face center localization that is approximated by using the object tracker. The mean absolute errors (MAE) were reduced from 1.40 and 1.38 degrees to 0.37 and 0.48 degrees for the horizontal and vertical directions, respectively. Strictly speaking, the ratio of the decrease on

[2]Some additional markers were also used, but they were omitted in this study; their horizontal angles ($\pm$60 degrees) were extreme for some subjects to correctly look at them, or the eyes while looking those markers were almost closed in our camera setting.

TABLE I

LOCALIZATION ERRORS OF FACE, IRIS, AND GAZE POINT

Target		MAE [degrees] horizontal	vertical
Face center	(TLD [29] only)	1.40	1.38
	(TLD [29] + Co-FPV)	0.37	0.48
Iris center	(modified IIP [32])	4.93	6.32
	(modified IIP [32] + bias-shift)	3.95	3.44
Gaze point	(manual marker + manual iris)	2.21	2.45
	(manual marker + automatic iris)	3.10	2.67
	(Co-FPV)	3.69	3.15

TABLE II

CORRECT RECOGNITION RATES OF GAZEES AND SPEAKERS

Target	Correct rates [%]
Gazee (total)	91.4
(while wearer is speaking)	84.9
(while wearer is listening)	96.1
Speaker	80.1

the horizontal axes (74%) to that on the vertical axes (65%) should be noted, because Co-FPV refined the face center only along the horizontal axis in this study. Moreover, the angle errors were calculated from pixel errors with the angle to pixel scale of 2.1, which was obtained by assuming that all eyes are looking exactly straight ahead and using the estimated eyeball size in the in-cam images (27.5 pixels).

2) Iris center localization: The MAEs are 4.9 and 6.3 degrees for both axes. Overall, the errors are larger than those of face center localization. However, we observed that IIP tends to bias the results especially in the y-axis compared to manual localization. Accordingly, Table I also shows the bias-removed errors for better understanding of the performance. The resulting MAEs are 4.0 and 3.4 degrees.

3) Gazee and speaker recognition: Table II shows that the proposed framework yields high correct recognition rates for both gazee and speaker recognition. Although quite simple, our model identified the speakers accurately enough. Moreover, gazee recognition is more challenging while the wearer is speaking than while listening, because speaker's gaze frequently switched among the listeners.

4) Gaze point estimation: The last three rows in Table I show errors in gaze point estimation obtained by comparing the estimated gaze points with the physical marker locations for three methods: a) both iris positions $\boldsymbol{e}$ and marker positions, corresponding to $\boldsymbol{g}$, were manually localized, b) the markers were manually determined but the irises were automatically localized, and c) both were automatically determined for the conversation data[3]. Co-FPV (c) yields comparable performance to marker-based calibration for both conversation sessions. This suggests that the differences in conversation types and the iris localization bias were well compensated in the mapping function training. Moreover, the MAEs of around 3.5 degrees and the average face width/height of 2.5-3.0 degrees (=40-50 pixels) suggest that though it is difficult to discriminate facial parts, e.g. eyes from mouth, it is possible to determine whether the wearer looked at face or body.

[3]We observed that the eye position aligned with a temporal filter was slightly but occasionally biased between the conversation sessions and marker-based calibration, due to differences in the spatial distribution of gaze direction. Accordingly, the difference in the mean was manually corrected for the marker-based calibration data to focus on the errors caused by the sample selection proposal; accurate helmet slip compensation is not a key focus of this paper.

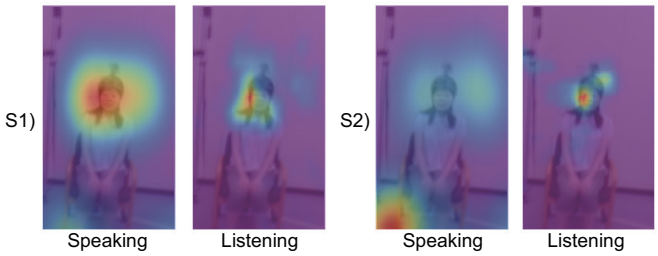

Fig. 11. Gazee-centered heat maps obtained from two subjects

F. Gaze heat maps

Fig. 11 shows heat maps from the first conversation session. They are separated into those obtained under speaking and listening conditions to clarify the interpersonal differences[4]. S1 is a typical subject who yielded similar gaze patterns in speaking and listening for both sessions. On the other hand, S2 was, while speaking, mainly looking at others' bodies, i.e. she appeared to avoid eye contact with the listeners. This tendency is expected to originate from her personality and emotional states, and/or social pressure, as previously suggested, e.g. [35].

VI. DISCUSSION

The experiment demonstrated the basic validity of Co-FPV. However, the current method has several issues.

1) Usability: Although our camera configuration, in its current form, does not seem usable in the wild, it is useful in the laboratory; it also has the potential to measure other nonverbal behaviors, such as facial expressions and head gestures, which are often required for emotion analysis.

2) Applicability: Among the assumptions made in this study, only the following two are crucial: people often converse with each other, and they mainly look at the face of the speaker. They are mostly true for a variety of conversation scenarios. However, they would be violated for people with autism, who tend to gaze at the other's body [21]. Some of the remaining constraints, e.g. standing or in-motion conversations, can be relaxed by introducing crowd tracking techniques, especially tracking-by-detection with data association, e.g. [36]; glass-type camera devices; and manual camera (internal parameter) calibration.

3) Conversation settings: The present study focused on medium-sized-party (group) conversations with the interlocutors sitting in a circle. Accordingly, it should be evaluated how the number of interlocutors, their spatial arrangement,

[4]Note that it is natural that the gaze point is mainly centered on the face in listening; it matches the constraint that we imposed on the training samples.

and changes in both over time impact the proposed framework. For example, if group size N decreases, the spatial distribution of others' faces should become more sparse; this suggest that training the eye-gaze mapping function would be difficult. However, on the other hand, the gazee of listeners would be more likely to be the speaker, since the chance level is $N-1$. This enables comparison of the proposed method with previous studies targeting small-sized conversations.

4) Gaze model: This paper dealt with gaze point as a two-dimensional point on the image, and used a simplified camera and geometry model. Although the experiment demonstrated that the approach works robustly in the conversation settings described, full three-dimensional modeling, as in [37], would be required to increase the accuracy.

5) Heat maps: Although this paper presented gaze duration heat maps by focusing on fixations, other visualizations are possible: e.g. other fixation-derived metrics, and saccade- and scanpath-derived metrics [38]. Additionally, pair-wise metrics, e.g. gaze following and eye contact, or group-wise metrics would be notable in social interactions. Determining the best metrics is another issue.

VII. Conclusions and Future Directions

We introduced the Co-FPV framework for automatic conversation analysis. Its eye-gaze mapping function is self-trained off-line by automatically selecting training samples based on conversational rules. Speaker diarization and the head tracking of interlocutors are performed by centralizing and processing the videos captured from each camera. This estimation approach yields a gazee-centered heat map for each interlocutor. An experiment using two six-person groups demonstrated the potential of the proposed framework.

This paper introduced FPV as a tool for automatically assessing nonverbal behaviors. However, FPV can also be an essential tool for obtaining emotion data. Our assumption is that first-person view images make it easier for the subject (or external observers) to recall (or read) his/her emotion felt at that time accurately, because the first-person perspective (or perspective taking for the observers) plays an important role in these processes [39], [40]. The evaluation of this assumption is a future task.

References

[1] D. Gatica-Perez, "Analyzing group interactions in conversations: A review," in *Proc. IEEE Int'l Conf. MFI*, 2006, pp. 41–46.

[2] K. Otsuka, "Conversation scene analysis," *IEEE Signal Proc. Mag.*, vol. 28, pp. 127–131, 2011.

[3] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vision Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[4] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *EmoSPACE*, 2013, pp. 1–8.

[5] C.-H. Wua, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA Trans. Sig. Info. Proc.*, vol. 44, no. 3, pp. 572–587, 2011.

[6] G. Mckeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schroeder, "The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent," *IEEE TAC*, vol. 3, pp. 5–17, April 2012, issue 1.

[7] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (CERT)," in *FG*, 2011, pp. 298–305.

[8] R. El Kaliouby, R. Picard, and S. Baron-Cohen, "Affective computing and autism," *Ann. N. Y. Acad. Sci.*, vol. 1093, no. 1, pp. 228–248, 2006.

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

[10] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato, "Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings," in *Proc. IEEE Int'l Conf. FG'11*, 2011, pp. 43–50.

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

[12] R. Vertegaal, "Look who's talking to whom," *Ph.D. thesis, University of Twente*, 1998.

[13] K. Otsuka, Y. Takemae, and J. Yamato, "A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances," in *ICMI*, 2005, pp. 191–198.

[14] R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe, "On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions," in *Proc. ICMI*, 2013, pp. 3–10.

[15] M. Taj and A. Cavallaro, "Distributed and decentralized multicamera tracking," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 46–58, 2011.

[16] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *J Vis*, vol. 14, no. 8 article 5, pp. 1–17, 2014.

[17] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *arXiv:1409.1484*, 2014.

[18] T. Kanade and M. Hebert, "First-person vision," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2442–2453, 2012.

[19] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *CVPR*, 2012, pp. 1226–1233.

[20] F. Martinez, A. Carbone, and E. Pissaloux, "Combining first-person and third-person gaze for attention recognition," in *FG*, 2013, pp. 1–6.

[21] L. Speer, A. Cook, W. McMahon, and E. Clark, "Face processing in children with autism," *Autism*, vol. 11, no. 3, pp. 265–277, 2007.

[22] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis Res*, vol. 40, pp. 1489–1506, 2000.

[23] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze." *PAMI*, vol. 32, no. 3, pp. 478–500, 2010.

[24] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *PAMI*, vol. 35, no. 2, pp. 329–341, 2013.

[25] F. Alnajar, T. Gevers, R. Valenti, and S. Ghebreab, "Calibration-free gaze estimation using human gaze patterns," in *ICCV*, 2013.

[26] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, "Pose-invariant facial expression recognition using variable-intensity templates," *IJCV*, vol. 83, pp. 178–194, 2009.

[27] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. ETRA*, 2000, pp. 71–78.

[28] A. Bojko, "Informative or misleading? heatmaps deconstructed," in *Proc. HCII*, 2009, pp. 30–39.

[29] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *PAMI*, vol. 34, no. 7, pp. 1409–1422, 2012.

[30] C. Zhang and Z. Zhango, "A survey of recent advances in face detection," *Microsoft Research Tech. Rep., MSR-TR-2010-66*, 2010.

[31] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001, pp. 511–518.

[32] R. Valenti and T. Gevers, "Accurate eye center location through invariant isocentric patterns," *PAMI*, vol. 34, no. 9, pp. 1785–1798, 2012.

[33] J. Schwiegerling, *Field Guide to Visual and Opthalmic Optics ,.* SPIE Press, 2004.

[34] M. Chen, "Leveraging the asymmetric sensitivity of eye contact for videoconference," in *CHI*, 2002, pp. 49–56.

[35] R. J. Larsen and T. K. Shackelford, "Gaze avoidance: personality and social judgments of people who avoid direct face-to-face contact," *Pers. Indiv. Differ.*, vol. 21, no. 6, pp. 907–917, 1996.

[36] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE TPAMI*, vol. 36, no. 1, pp. 58–72, 2014.

[37] H. S. Park, E. Jain, and Y. Sheikh, "3d social saliency from head-mounted cameras," in *NIPS*, 2012, pp. 431–439.
[38] A. Poole and L. J. Ball, *Encyclopedia of Human Computer Interaction*. Pennsylvania: Idea Group, 2004, ch. Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects.
[39] A. D'Argembeau, C. Comblain, and M. van der Linden, "Phenomenal characteristics of autobiographical memories for positive, negative, and neutral events," *Appl. Cog. Psychol.*, vol. 17, no. 3, pp. 281–294, 2003.
[40] M. H. Davis, "Measuring individual differences in empathy: Evidence for a multidimensional approach," *J. Pers. Soc. Psychol.*, vol. 44, no. 1, pp. 113–126, 1983.