Collective First-Person Vision for Automatic Gaze Analysis in Multiparty Conversations

Shiro Kumano, *Member, IEEE*, Kazuhiro Otsuka, *Member, IEEE*, Ryo Ishii,
and Junji Yamato, *Senior Member, IEEE*

***Abstract*—This paper targets small- to medium-sized-group face-to-face conversations where each person wears a dual-view camera, consisting of inward- and outward-looking cameras, and presents an almost fully automatic but accurate off-line gaze analysis framework that does not require users to perform any calibration steps. Our collective first-person vision (Co-FPV) framework, where captured audio-visual signals are gathered and processed in a centralized system, accurately and jointly undertakes the fundamental functions required for group gaze analysis, including speaker detection, face tracking, and gaze tracking. Of particular note is our self-calibration of gaze trackers by exploiting a general conversation rule, namely that listeners are likely to look at the speaker. From the rough conversational prior knowledge, our system visualizes fine grained participants' gaze behavior as a gazee-centered heat map, which quantitatively reveals what parts of the gazee's body the participant looked at and for how long while the gazer was speaking or listening. An experiment using conversations amounting to a total of 140 min, each lasting an average of 8.7 min and engaged in by 37 participants in groups of three to six, achieves a mean absolute error of 2.8 degrees in gaze tracking. A statistical test reveals neither a group size effect nor a conversation type effect. Our method achieves F-scores of over .89 and .87 in gazee and mutual gaze recognition, respectively, in comparison with human annotation.**

***Index Terms*—First-person vision, egocentric vision, wearable camera, conversation, gaze, mutual gaze, eye contact, self-calibration, head pose estimation, speaker diarization**

I. Introduction

FACE-to-face conversation is the primary way by which we share information, understand others' skills, personalities and emotions, and make decisions in daily life, e.g. meetings or job interviews. It is a highly multimodal process that mainly involves the use of audio and visual signals to perceive/understand others' verbal/nonverbal behaviors, such as an individual's speech, gaze, facial expressions, gestures and postures as well as his/her interactions with the group. In this regard, automatic multimodal conversation analysis is acknowledged by the multimedia research community as a basic research area in relation to the development of computer-mediated communication or conversational agents [1], [2]. Measuring these behaviors is a preliminary step toward inferring high-level states, such as personality traits [3], functional roles [4], hirability [5], affective states [6], and empathy [7].

Conversation is often captured by microphones and cameras (as well as depth sensors, as in [8], [9]) *installed in the environment*. However, these settings frequently require human annotation when conversational behaviors are investigated in detail. This is a particularly severe requirement for medium-sized (five to ten participants) or large conversations. For example, the directions of faces in relation to fixed cameras vary greatly when people look around at the other participants. In such scenarios, it is still hard for state-of-the-art passive-vision-based face/gaze analyzers, e.g. [10], [11], or even human coders [12] to fully distinguish subtle facial/gaze changes and head movements.

One of the most fundamental of these nonverbal behaviors is gaze given its importance in a number of social functions [13], including monitoring, visual feedback, expressing emotion, and regulating the flow of a conversation. Of particular note for this study is that people pay attention by orienting their gaze toward the speaker [14], [15]. However, even when using off-the-shelf glasses-type eye trackers, the collection of accurate group gaze behavior, e.g. who is looking at whom, and at which facial/body part, and when, is still time-consuming due to the need for manual intervention. It is for calibrating the gaze trackers by asking the wearer to fixate on some predefined points on a screen or in the environment, as in [10], [16]; localizing faces in the field of view (FOV) images or simply determining the ID of the gazing target [17]; and identifying the speaker at every moment, although speaker detection is often automated, as in [17]. The cost increases as the group size increases. Consequently, most previous conversation analyses targeted dialogues [14], [18], trialogues [15], [19], and quadlogues [20], [21], or use head poses in medium-sized social interactions as rough estimates of the visual focus-of-attention, e.g. [22], [23].

Our way of overcoming these barriers is to introduce first-person view or egocentric images thanks to the recent rapid industrial and algorithmic growth of this field [24]. This paper considers a situation, where the wearer's FOV is captured by an outward-facing camera (out-cam), along with his/her face captured by an inward-facing camera (in-cam) with microphones. These cameras, called dual-view

The authors are with the NTT Communication Science Laboratories, NTT Corporation, Kanagawa 243-0198, Japan (e-mail: kumano@ieee.org; kazuhiro.otsuka@lab.ntt.co.jp; ishii.ryo@lab.ntt.co.jp). Find the published version of this article under DOI: 10.1109/TMM.2016.2608002.

cameras here, are assumed to *be attached to a rigid worn item*, e.g. a helmet or glasses. First-person-view images are inherently superior for measuring certain behaviors of the wearer because the facial motions and head motions are mostly separated by these cameras. The wearer's face is nearly always stable in the in-cam image, as in [25], while any head motion yields a large motion flow in the out-cam image, and this is useful for head gesture recognition, e.g. [26].

Our basic idea is to unify the series of fundamental computer vision/audio signal processing techniques required for these analytical steps with the aim of developing an almost fully automatic system that offers deeper conversation analysis. We combine two established frameworks. The first is interpersonal sensor fusion, where signals captured by multiple sensors (everyone's cameras in our case) are collected in a centralized system, or shared among distributed systems, as in [27]. The second is the exploitation of the primary characteristic of multi-party conversation that has already been elucidated, namely that the listeners in conversation are likely to look at the speaker [13], [15], [28], as validated once again in Section V-C1. We call this framework, which subsumes the above two frameworks, *Collective First-Person Vision (Co-FPV)*. Figure 1 illustrates the proposed framework.

We show in Section III-B1 that Co-FPV via interpersonal sensor fusion makes the face-tracking problem very easy to deal with. Our key idea is to track a person's face from the out-cam image *captured by the person*. The conversation rule is used in Section III-A1 as prior knowledge with which to roughly calibrate everyone's gaze tracker. This is the initial stage preceding fine-tuning. Here we employ only samples where the wearer is estimated to be looking at the speaker's face with high likelihood. These techniques achieve the automatic characterization of the fine-grained gaze behavior of participants as a *gazee-centered* heat map, which reveals which part of the gazee the participant looked at, and for how long. We demonstrate that Co-FPV performs promisingly despite its simplicity by undertaking an experiment involving a total of 140 minutes of conversations engaged by 37 participants in groups of three to six.

The contributions of this paper are as follows: 1) Co-FPV, which unifies both the required techniques and everyone's observations, 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 interpersonal sensor fusion in the Co-FPV framework is proposed.

The remainder of this paper is organized as follows. Section II describes related work. Section III explains the proposed framework. Sections IV, V and VI detail the experiment, the results, and the corresponding discussion. Section VII provides a summary and outlines future work.

II. Related Work

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

A. Fixed-camera-based gaze analysis in conversation

The automatic recognition of the visual focus of attention in a conversation has been tackled mainly by using cameras fixed in the environment, as reported in [18], [20], [21], [29]. This contactless approach is advantageous in terms of usability, but often suffers from the participant head rotation mentioned above. Some researchers have tried to exploit other behaviors of the target person as additional information, including the head pose and audio signals. However, the performance is still limited, and this makes it difficult to handle irregular cases where the assumed combination of gaze and these behaviors is broken.

B. First-person vision (egocentric vision)

FPV is a hot topic in the computer vision community, and the number of related papers is increasing rapidly [30]. Several tasks have already been tackled: including gaze tracking, activity recognition, and video summarization [24]. Of particular note is the pioneering work that targets social interactions in FPV. In [31], the type of social interaction, i.e. monologue, dialogue or discussion, is classified from the poses of people in images captured by a camera that is being worn. A similar task is tackled in [19] by combining egocentric images with images taken by a stationery camera installed in the environment. In [32], social saliency, which is defined in terms of a spot that many people are likely to look at, is predicted by a single wearable camera. However, these studies do not consider situations where everyone has his/her own wearable camera(s). Arev et al. [22] recently focused on such a situation, they approximated each person's gaze direction with the head pose orientation. Although this approach provides a sufficient approximation for video editing from multiple egocentric videos, it is hard to fully detect side-glances without using any information about the eyes.

C. Gaze analysis with eye tracker

The use of eye trackers to understand how people observe photographs or movies has long been a subject of research. A good example involves investigating the difference between the gaze behaviors of typically developed people and people with autism, as described in [33]. Another topic involves building a computational/stochastic model that well explains human gaze patterns. Various models have already been proposed, such as Itti and Koch's model [34] of low-level saliency, and a high-level objective model. For example, it is reported in [28] that while observing a movie depicting social interaction, people tend to look at the turn holder, and low-level saliency models fail to

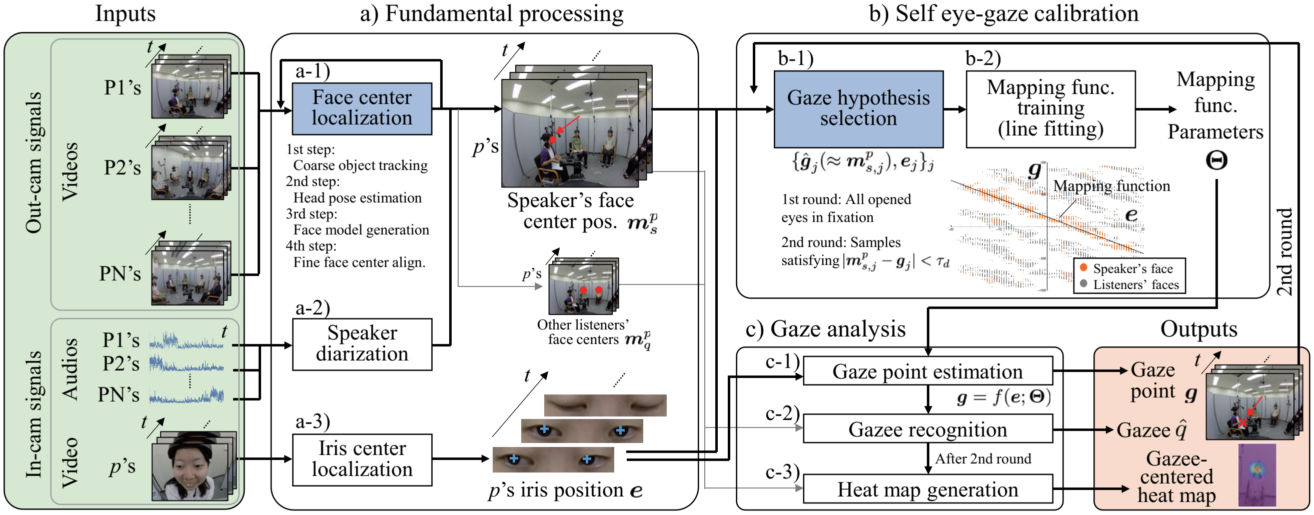

Inputs
Out-cam signals
Videos
P1's
P2's
PN's
In-cam signals
Audios
P1's
P2's
PN's
Video
p's
t
a) Fundamental processing
a-1)
Face center localization
1st step: Coarse object tracking
2nd step: Head pose estimation
3rd step: Face model generation
4th step: Fine face center align.
p's
Speaker's face center pos. $\boldsymbol{m}_s^p$
p's
Other listeners' face centers $\boldsymbol{m}_q^p$
a-2)
Speaker diarization
a-3)
Iris center localization
p's iris position $\boldsymbol{e}$
b) Self eye-gaze calibration
b-1)
Gaze hypothesis selection
$\{\hat{\boldsymbol{g}}_j(\approx \boldsymbol{m}_{s,j}^p), \boldsymbol{e}_j\}_j$
1st round: All opened eyes in fixation
2nd round: Samples satisfying $|\boldsymbol{m}_{s,j}^p - \boldsymbol{g}_j| < \tau_d$
b-2)
Mapping func. training (line fitting)
Mapping func. Parameters $\boldsymbol{\Theta}$
$\boldsymbol{g}$
Mapping function
$\boldsymbol{e}$
● Speaker's face
● Listeners' faces
2nd round
c) Gaze analysis
c-1)
Gaze point estimation
$\boldsymbol{g} = f(\boldsymbol{e}; \boldsymbol{\Theta})$
c-2)
Gazee recognition
After 2nd round
c-3)
Heat map generation
Outputs
Gaze point $\boldsymbol{g}$
Gazee $\hat{q}$
Gazee-centered heat map

Fig. 1. Overview of the proposed Co-FPV gaze analysis (the flow for wearer p only): Our key contributions are highlighted in blue. a) The first step is fundamental processes: face localization in p's out-cam images, speaker diarization from in-cam audio signals, and iris center localization in p's in-cam images. The face positions (and orientations) and speaker are inferred by using all group members' signals (i.e. Co-FPV approach), while the iris centers are localized by using only p's own in-cam images (i.e. a non-Co-FPV approach). b) By using sets of the speaker's face positions $\boldsymbol{m}_s^p$ and the corresponding iris positions $\boldsymbol{e}$ as hypotheses, the parameters of the eye-gaze mapping function f (i.e. $f : \boldsymbol{e} \mapsto \boldsymbol{m}_s^p$), $\boldsymbol{\Theta}$, is trained. c) After self-calibration, p's gaze point $\boldsymbol{g}$ at each moment is estimated by substituting $\boldsymbol{e}$ into f, and the gazee is recognized as the nearest person to the gaze point. Finally, p's gazee-centered heat map is generated by using the gaze point and the gazee's face coordinate system (i.e. the gazee's face location and orientation) at each moment.

explain such gaze patterns. Gaze trackers are also used for automatic meeting analysis, e.g. [17]. 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 all these processes.

D. Self-calibration of eye-gaze mapping function

Although there are numerous gaze tracking techniques [35], only a few self-calibration techniques for passive vision-based eye trackers have been proposed. These studies are based on stochastic prediction of the gaze point by using a low-level visual saliency model [36], [37], [38], the coordination of eye, head and hand movements in object manipulation tasks [39], or use others' gaze patterns to target images [40]. However, the first two are not expected to work well when viewing multi-party social interaction [28], and the last is inapplicable to unknown conversational scenes. The current work differs from those studies mainly in the sense that conversational saliency, namely the dominant visual focus of attention of the turn holder, is used as prior knowledge with which to predict the gaze point of a target person without using the gaze behavior of others[1].

III. Proposed method

This section describes our proposed automatic gaze analysis for Co-FPV. Our main task is to estimate, without manual intervention, where each camera-wearer looked at at each moment during group conversation. We always consider in this section that the target wearer is $p \in \mathcal{Q}$, where $\mathcal{Q} = \{1, \cdots, N\}$, and N is the number of people in the conversation. Hereafter symbol p is omitted, unless necessary. The gaze point is defined as two-dimensional coordinates in p's out-cam image, $\boldsymbol{g} = (g_x, g_y)^{\mathrm{T}}$. For this purpose, we incorporate the iris center coordinates localized in p's in-cam image at that moment, $\boldsymbol{e} = (e_x, e_y)^{\mathrm{T}}$. The task is now replaced with the estimation of a bijective mapping function, f, that associates iris center coordinates $\boldsymbol{e}$ with the corresponding gaze point $\boldsymbol{g}$, namely $\boldsymbol{g} = f(\boldsymbol{e}; \Theta)$, where Θ denotes the parameters of f. Moreover, the origins of $\boldsymbol{g}$ and $\boldsymbol{e}$ are the image centers.

The flow of our system is as follows and is also illustrated in Fig. 1. The headings in the manuscript correspond to those in the figure. Our system is mainly characterized by its self-training of parameters Θ in multi-person conversation scenarios by assuming that the listener is looking at the center of the speaker's face, defined as the mid-point between the eyes. Our processes, described below, mostly run in parallel for all the conversation participants, although some run for the entire group.

a) Fundamental processing: We receive all the participants' in- and out-cam signals as input. First, we localize the speaker's face in p's out-cam image, $\boldsymbol{m}_s^p$, and iris center in p's in-cam image $\boldsymbol{e}$ at each moment. The speaker's face localization consists of speaker diarization and face center localization. a-1) Everyone's face in p's out-cam is

[1] We have already proposed our research framework in [41]. This paper provides; a more sophisticated two-stage self-calibration procedure (only the first stage was proposed in [41]); an extended experiment, including various conversation group sizes (different numbers of participants) and spatial arrangements; more comprehensive theoretical bases of the proposed method, including a more rigorous hypothesis testing our basic assumption; improved iris center detection; a more substantial survey of related work; and a comparison with state-of-the-art methods.

accurately localized with her ID by using a two-dimensional object tracking technique and the proposed Co-FPV head pose estimation (Section III-B1). A notable difference between our Co-FPV method and the traditional methods is that we use not only p's own out-cam images but also the others' out-cam images to improve efficacy and accuracy. a-2) The speaker s is identified as the person generating the maximum acoustic power in everybody's in-cams at any given moment (Section III-C3). As a result of the above processes, $\boldsymbol{m}_s^p$ is obtained. a-3) Iris center $\boldsymbol{e}$ is localized in p's in-cam images by using an existing technique [42] (Section III-C4).

b) The eye-gaze mapping function f is then self-trained by assuming that the listener is always looking at the speaker's face, i.e. $\boldsymbol{g} \approx \boldsymbol{m}_s^p$ (Section III-A1). This is performed in two steps. b-1) All the samples are used in the first stage, but false samples, detected by using the estimated gaze point (explained next), are ignored in the second stage. b-2) Parameters Θ of f are then trained as those minimizing the distance between the resulting $\boldsymbol{g}$ and $\boldsymbol{m}_s^p$.

c) Gaze analysis: c-1) The gaze point of each person in the out-cam images is then estimated by substituting the iris center $\boldsymbol{e}$ into the trained mapping function f. c-2) The gazee is recognized as the person nearest to the gaze point. c-3) Finally, a gazee-centered heat map for each wearer is generated through gazee recognition (Section III-A2).

The following subsections describe these processes. Note that Sections III-A1 and III-B contain our key proposal, while the other subsections describe fundamental techniques, which have less novelty or can be easily replaced with alternatives.

A. Proposed automatic gaze analysis framework

1) Self-calibration of eye-gaze mapping function: Parameters Θ of the mapping function f are obtained by minimizing the following objective function:

$$\Theta = \arg\min_{\Theta} \Sigma_j dist(f(\boldsymbol{e}_j;\Theta), \tilde{\boldsymbol{g}}_j), \tag{1}$$

where $dist$ is a distance function, and j denotes a sample index in a sample set, $\{\tilde{\boldsymbol{g}}_j, \boldsymbol{e}_j\}_j$. The concrete form of mapping function f used in this study is given in Section III-C1. We used the Random Sample Consensus (RANSAC) [43], which is a frequently used robust estimation technique. One of the main advantages of the RANSAC algorithm is that it requires no detailed prior knowledge about parameters.

We propose a two-stage coarse-to-fine self-calibration to solve Eq. 1; that is, we use different training sets in the first and second stages. The first set, by assuming that the gaze point is the face center of speaker s at that moment, i.e. $\tilde{\boldsymbol{g}} = \boldsymbol{m}_s^p$, consists of all samples that satisfy the following two conditions: the face in the out-cam is the speaker, and the wearer's gaze is in a state of fixation. Fixations are detected based on the dispersion as a short temporal block with a small variation in $\boldsymbol{e}$ without blinking or eye closure; this is the dispersion-threshold identification [44]. These constraints are not perfect and sometimes yield outliers. For example, in some cases the wearer may look at another listener.

To further eliminate such outliers, we, in the second stage, update the parameters solely by using the samples where the wearer is, after the first stage, estimated to be really looking at the speaker's face according to the procedure explained next.

2) Gaze analysis (gaze point estimation and heat map generation): After the mapping function parameters have been learned, the gaze point at each time $\boldsymbol{g}$ is obtained by substituting the iris center coordinates $\boldsymbol{e}$ at that time into f. The gazee, $\hat{q} \in \mathcal{Q}\backslash p$ (where $\mathcal{Q}\backslash p = \{1, \cdots, p-1, p+1, \cdots, N\}$), is recognized as the person nearest to the gaze point, as determined by Euclidean distance, namely the task is a multi-class classification task. If the distance exceeds threshold τ_d, the wearer is not considered to be looking at anyone's face. After the gazee has been determined for each participant, namely mutual gaze or eye contact, whether a pair of people are looking at each other or not, is identified. Although this is a binary classification task, it is more challenging than the individual's gazee recognition, because both mutual gazers must be correctly recognized.

We then generate a gazee-centered heat map as a (relative) gaze duration heat map [45], which shows the accumulated time the wearer spent looking at the different areas of the other interlocutors. The estimated gaze point $\boldsymbol{g}$ in the gazer's out-cam coordinates is mapped into gazee $\hat{q}$'s face coordinate system as: $\boldsymbol{\xi} = 2/w_{\hat{q}} \cdot (\boldsymbol{g} - \boldsymbol{m}_{\hat{q}})$, where $w_{\hat{q}}$ is the width of $\hat{q}$'s facial skin bounding box. Heat maps are then obtained by collecting $\boldsymbol{\xi}$ values during gaze fixation and visualizing their density.

B. Proposed fine face localization in out-cam images

The face center location of each person in the out-cam image of another $\boldsymbol{m}$ at each moment is required for self-calibration in Section III-A1 ($\boldsymbol{m}_s^p$), head pose estimation in Section III-B1 ($\boldsymbol{m}_p^q$ for all $q \in \mathcal{Q}\backslash p$), gazee recognition in Section III-A2 ($\boldsymbol{m}_q^p$), and gazee-centered heat map generation in Section III-A2 ($\boldsymbol{m}_{\hat{q}}^p$). This subsection explains how we obtain the face center locations.

We assume that the face location of each person, $\tilde{\boldsymbol{m}}_{\cdot}^{\cdot}$, has already been roughly estimated by using an object tracking technique [46] (described in Section III-C2) in the first step; e.g. $\tilde{\boldsymbol{m}}_{\boldsymbol{p}}^{\boldsymbol{q}}$ represents p's rough face center coordinates in q's out-cam image. The aim of the second step is to refine the face center position. We achieve this by rotating a three-dimensional frontal face model around x (horizontal), y (vertical) and z (in-plane) axes according to the head pose, $\boldsymbol{\phi} = (\phi_x, \phi_y, \phi_z)$, accurately estimated by our method. Figure 4 shows the flow of the proposed face localization.

1) Head pose estimation: In most previous FPV studies, the head orientation of person q is estimated from the out-cam image of *the target wearer* (i.e. p), e.g. [31], relying on the change in q's facial appearance when observed from different relative view orientations. On the other hand, we propose obtaining person q's head pose from p's face position in q's *own* out-cam image (i.e. $\boldsymbol{m}_p^q$); for this we use geometrical optics.

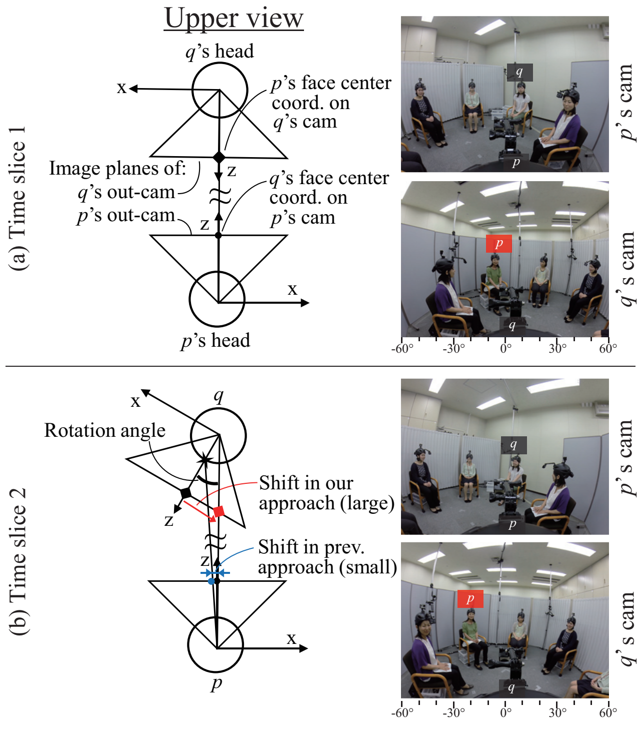


Fig. 2. Relation between facial pose and position in out-cam images: p's face positions in q's images (lower images in both time slices), i.e. $\boldsymbol{m}_p^q$, provide the face poses of q in p's images (upper images in both time slices). In this case, the horizontal angles of q relative to p in both time slices are close to 0°and -20°, respectively. Compare the horizontal positions of red boxes on the right hand side.

Previous approaches, e.g. [31], are basically very similar to traditional fixed-camera-based approaches, which have been well surveyed in the literature, e.g. [47]. In short, they often rely on the change in relative spatial arrangement of facial landmarks or small patches caused by head rotation.

We, on the other hand, use the fact that the geometrical optics of typical lenses provide a one-to-one (roughly linear) correspondence between q's head pose in relation to p, $\boldsymbol{\phi}$, and p's position in q's out-cam image. For mathematical simplicity, we here assume that everyone's face and out-cam are facing in the same direction (although considered in the mapping function, as in Section III-C1) and the locations of their centers are the same, the images are captured through the equidistance projection [48] (which is frequently used for wide-angle camera perspectives, e.g. [49]), and no one rotates their head in the in-plane direction[2], i.e. $\phi_z = 0$. Now, the correspondence can be approximated as $(\phi_y, \phi_x)^{\mathrm{T}} = \tilde{\boldsymbol{m}}_p^q/\alpha_o$, where α_o is the out-cam's scale factor that relates degrees to pixels in both axes[3], namely the ratio between the number of pixels and the degrees of the camera's FOV. For example, if p is located in the middle (or on the left side) 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. 2.

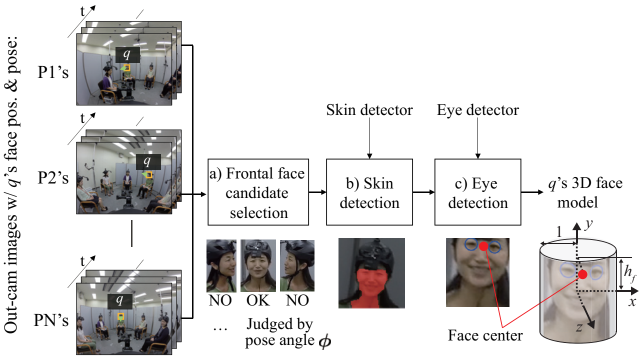


Fig. 3. Flow of our face model generation for person q: a) Frontal face images of q are collected continuously from the images of the others' out-cams. b) The skin regions (red masks), and c) the centroid of both eyes, i.e. face center, shown as a red dot, are detected in the collected frontal faces, although only a single image is shown here. d) Their mean vertical position normalized by the face size is defined as h_f in the model.

We can theoretically compare our approach with the traditional approach by considering an example case where the front of q's face, with a width of 40 pixels in p's image, is rotated horizontally by one degree, i.e. $\phi_y = 1°$. If the camera horizontal FOV is 1920 pixels and 122.6° (i.e. $\alpha_o = 1920/122.6 = 16$), then p's face center shifts by $m_{p,x}^q = \alpha_o \cdot 1° = 16$ pixels in q's image (proposed approach), as shown by the red arrow in the lower left of Fig. 2; while q's face center moves by only 0.35 (=$40/2 \cdot \sin(1°)$) pixels in p's image (traditional approach), as shown by the short blue line sandwiched between the blue arrows in the lower-left of Fig. 2. Even when we consider the face alignment error (8 pixels = $\alpha_o \cdot 0.43°$, as later assessed in Section V-A2), the shift in our approach is much larger than the shift in the traditional approach.

2) Face model: The face model of each person consists of a three-dimensional shape and the position of the face center is indicated 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 must be almost frontal; this is judged by using $\boldsymbol{\phi}$, and is estimated in Section III-B1. Second, both eyes are detected in the image with 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. Moreover, the face model of each person is shared among the group members. The frontal face condition often forced each person's face to be captured by the person who was sitting in front of her. This is because, for example, the person between p and q, wearing a purple cardigan, in Fig. 2 rarely turns her head completely toward p or q when changing gaze direction [20].

We use a cylinder as a face shape model, as in [51]. The face center is defined on its surface. The directions of the face coordinate system of the axes are horizontal, vertical and facial-frontal. The face center coordinate is $\boldsymbol{x}_c = (0, h_f, 1)^{\mathrm{T}}$. The height h_f is the mean y-coordinate of the mid-point between the two eyes in the bounding

[2] This rough approximation worked well in our experiment, as demonstrated in Section V-A2. However, if large roll rotations occur frequently, a sophisticated motion estimation algorithm, e.g. [50], would be required.

[3] If the scale factor is different for both axes, then ϕ_x and ϕ_y should be obtained with different α_os.

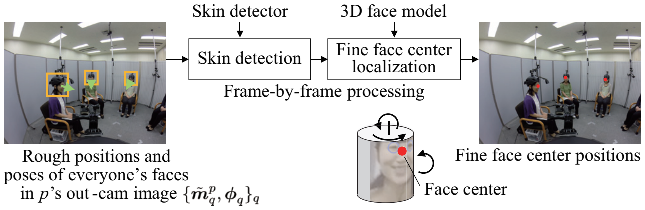

Skin detector
3D face model
Skin detection
Fine face center localization
Frame-by-frame processing
Rough positions and poses of everyone's faces in p's out-cam image $\{\tilde{\boldsymbol{m}}_q^p, \boldsymbol{\phi}_q\}_q$
Fine face center positions
Face center

Fig. 4. Flow of our fine face center localization from coarse face positions obtained by object tracker [46] and fine rotation angles obtained with the proposed method, described in Section III-B: The face center coordinates of person q in p's out-cam image $\boldsymbol{m}_q^p$ are determined through detecting skin pixels around the rough face region and scaling and rotating q's face model, where the location of the face center is defined. The scaling factor is determined by using the skin region.

boxes of the skin region divided by the half width of the boxes. The skin region is detected by color thresholding-based masking in the hue, saturation and value (HSV) space [52] around the rough estimate of the face position obtained in the first step. Moreover, more sophisticated skin detection methods, e.g. found in [53], can be applied, if necessary[4]. Figure 3 shows the flow of our face model generation.

3) Fine face localization: The face center coordinate in an out-cam is refined as

$$\boldsymbol{m} = w f_c(\boldsymbol{R}_{\boldsymbol{\phi}} \boldsymbol{x}_c) + \bar{\boldsymbol{m}}, \tag{2}$$

where w represents the half width of the skin bounding box of the target person in the current out-cam image, $\boldsymbol{R}_{\boldsymbol{\phi}}$ denotes a three-dimensional rotation matrix for $\boldsymbol{\phi}$, and $\bar{\boldsymbol{m}}$ is the image coordinate of the center of the skin region. We here assume, like the orthographic projection, that the depth of the skin region along the line of sight is small compared to the distance from p's out-cam, i.e. $f_c(\boldsymbol{x}) = (x, y)^{\mathrm{T}}$, where $\boldsymbol{x} = (x, y, z)^{\mathrm{T}}$. Figure 4 shows the flow.

C. Existing models and techniques

This subsection describes the remaining models and techniques, required for the proposed method. Note that they are not the main part of our proposal and are replaceable.

1) Mapping function f: The exact form of the eye-gaze mapping function f is complex and nonlinear due the three-dimensional geometry of the camera optics and eyeball rotations. To focus on evaluating the self-calibration proposed in Section III-A1, we use a practical two-dimensional linear model, called a similarity transformation; i.e. we assume a linear relationship between the iris centers in the in-cam and the corresponding gaze point in the out-cam as

$$f(\boldsymbol{e}; \boldsymbol{\Theta}) \approx \begin{pmatrix} -a & 0 & b_x \\ 0 & a & b_y \end{pmatrix} \boldsymbol{e}', \tag{3}$$

where $\boldsymbol{e}'$ denotes the augmented vector of $\boldsymbol{e}$. In this model, there are three parameters: $\Theta = (a, b_x, b_y)^{\mathrm{T}}$. Parameter a indicates the slope of the linear function (identical for both axes), and parameters b_x and b_y are their intercepts. See Appendix for the derivation of Eq. 3.

The linear approximation introduces large errors when the gaze direction exceeds roughly 60°. However, such cases are infrequent with our settings, and we demonstrate in Section V-C2 that this linearization is sufficient. If the camera is very close to the face and has significant radial distortion, e.g. for glasses-type devices, the more exact form introduced in Appendix would be superior. Furthermore, we simply consider the centroid of both eyes as $\boldsymbol{e}$ in Eq. 3, although Eq. 3 can be trained for both eyes independently. This hampers gaze estimation in the depth direction but is practical and beneficial for avoiding overfitting.

2) Coarse head tracking: The tracking-learning-detection (TLD) tracker [46] was used to obtain the coarse location of faces in the out-cam images $\tilde{\boldsymbol{m}}^{\cdot}_{\cdot}$. 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. Initialization is the only manual intervention needed for the proposed method, and this can be automated by employing the multi-view face detector, introduced in [54], and person identification. The face captured by the in-cam would be used as the person's face model.

3) Speaker diarization: We defined the speaker s as the person generating the maximum acoustic power v in everybody's in-cams at any given moment. If the maximum power is lower than the threshold τ_v, no one is assumed to be speaking. The speaker diarization is written as:

$$s = \begin{cases} \arg\max_p v_p, & \text{if } v_s \geq \tau_v \\ \emptyset, & \text{otherwise,} \end{cases} \tag{4}$$

where $\emptyset$ means no one is speaking. This ignores simultaneous utterances[5].

4) Iris center localization: The in-cam images were aligned in advance to ensure that the horizontal coordinates of both eyes are the same and their centroid is located at the center of the image. This eliminates occasional slight helmets shift during conversation. The alignment was accomplished in the following steps: first, both eyes were localized by an eye detector [55]. Then, the eye positions were smoothed with a temporal filter to compensate for errors and missing observations. Finally, the resulting translation vector and rotation matrix were applied to the images.

To localize the iris centers in the eye-aligned images, we used the Fast Radial Symmetry Transform (FRST) [42] for both eyes. We made several modifications to the FRST. i) L_0 smoothing [56] was applied to the input images as preprocessing to eliminate noisy edges, such as

[4] Because all the participants in this study were Japanese, as explained in Section IV-A, we predefined the thresholds empirically. If the group consists of various ethnicities, it would be possible to train the skin detector individually by using faces detected in each participant's own in-cam images.

[5] We also tried using a person-wise binary threshold to determine whether a person was speaking or silent. This increased the recall but reduced the precision. However, as discussed in Section V-A1, precision is a more important factor for our framework. Thus, we chose the maximum-power-based method.

eyelashes, while preserving the iris edges. ii) We adopted a coarse-to-fine strategy; i.e. we ran the FRST twice. The results in the first stage were fed into the second stage as initial gaze directions. The initial gaze directions were used to selectively ignore unreliable edges with reference to [16]. For example, if the gaze was assumed to be leftward in the first stage, only the edges on the right hand side were used to determine the final results. We call it orientation-controlled FRST. Blinking and eye closures were determined by thresholding the vertical coordinate of the upper eyelashes detected as a dark region.

IV. Experimental settings

This section describes the experimental settings that we used to evaluate the performance of the proposed method. Our framework presented in Section III is so general that various hardware implementations (and the corresponding mapping functions) could be applied. Thus, the main aim of this experiment was to evaluate the validity of the principle of the proposed framework mainly by focusing on group size and conversation type. Designing smart hardware is beyond the main scope of this paper.

A. *Conversation dataset*

This paper targets various small- to medium-sized-group ($N = 3 - 6$) conversations. Thirty-seven Japanese women in their twenties to forties participated in this experiment. They were divided into non-overlapping groups consisting of three three-person groups, four four-person groups, and two six-person groups. Here, "M N-person groups" means that there are M groups, in each of which N people are interacting. Members of the same group had not met before the experiment.

The groups are categorized into four types according to their size and seating arrangement: G^3, G^4, $^*G^4$, and G^6. Here, the superscripts indicate the group sizes, and the asterisk denotes whether their seating arrangement is symmetric (no asterisk) or asymmetric (asterisk). Those in the symmetric (asymmetric) groups sat equidistant from each other in a circle (semicircle) with a radius of around 1.3 m. These seating arrangements are simply experimental setups, and our system automatically estimates the relative location of other participants for the target wearer via face center localization, as explained in Section III-B1.

All the symmetric groups engaged in two conversations. First, each member spent about 1.5 min introducing herself. The group members then participated in discussions and built a consensus as a group, that is they agreed on a single answer, related to a given topic within 10 min. Each of the asymmetric groups also held both self-introduction and discussion sessions, but only the discussion sessions were recorded. Moreover, the participants were not informed about the focus of this study and no instruction was given regarding gaze behavior. The spatial resolution of the cameras was mostly set at around full-HD, except for the in-cams for G^6. The temporal resolution of all the videos was set at 30 fps[6]. Table I summarizes our settings.

TABLE I
Summary of our conversation dataset

		Group types			
Properties		G^3	G^4	$^*G^4$	G^6
#Participants (N)		3	4	4	6
#Groups		3	2	2	2
Seat arrangement		Sym	Sym	Asym	Sym
In-cams:	Resolution	1080p	1080p	1080p	WVGA
	FOV	16x9W	16x9W	16x9W	16x9W
Out-cams:	Resolution	1080p	1080p	1080p	1440p
	FOV	16x9W	16x9W	16x9W	4x3W
Avg. conv. length [min]					
Self-introduction		4.7	6.0		7.9
Discussion		12.0	10.8	9.3	10.7
Marker-based calibration		✓	✓		✓

"1080p", "1440p" and "WVGA" indicate 1920 × 1080, 1920 × 1440 and 848 × 480 pixels, respectively. "16x9W" and "4x3W" denote 118.2° × 69.5° and 122.6° × 94.4°, respectively.

These cameras were synchronized by starting them simultaneously using a remote control. Moreover, in this work we did not eliminate the lens distortion of the cameras, namely we did not calibrate the cameras to obtain the intrinsic parameters[7]. The aim was to keep the original FOV angles, which were already close to the human horizontal visual FOV, more correctly the binocular field of view of 120° [57]. When the distortion is very severe, camera calibration would be required.

A crucial factor as regards the proposed framework, in addition to the precondition on the gaze behavior of interlocutors, i.e. the listeners are likely to look at the speaker, would be the turn-taking structure, i.e. whether they simply spoke in turn, as in the self-introduction sessions, or not, as in a heated discussion; we call these two types of conversations *ordered* and *unordered* conversations, respectively. In most discussion sessions, the group members spoke their own opinions in turn (34 ± 7 sec per subject) at the beginning (but not at the end), although they were not instructed to do so by the experimenters. That is, in terms of structure, such discussion sessions can be considered as a mixture of ordered (at the beginning) and unordered (subsequent) conversations.

We expected unordered conversations to be more challenging in our task, because the assumed conversational rule would be often violated. In ordered conversations, there is only a single speaker at any given moment, and the listeners are expected to be more likely to look at the speaker than in unordered conversations. However, in unordered conversations, overlapping speech would often occur as a result of listeners' backchannel behavior, and listeners occasionally look at other listeners to understand the conversation situation as regards obtaining a turn.

[6] To be precise, the frame rates of the cameras differed during recording, and they were re-rendered at 30 fps after the recording. In this step, we verified that there were no dropped frames and that the frames were synchronized.

[7] The radial distortion was insignificant in our experimental settings, because the faces were mainly located at the mid-level in the out-cam images.

B. Measurement device

Two cameras were attached to a lightweight mountain climbing helmet. The in-cam was mounted on a carbon shaft around 20 cm in front of the face without impeding eye contact with others, while the out-cam was placed at the top of the helmet. The helmet was counter-balanced to alleviate any shift during head motion. The spatial resolutions of the cameras, that is the numbers of pixels and FOV angles, were assumed to be already known, as summarized in Table I. The usability issues with this design are discussed in Section VI-A.

C. Parameter settings

Mapping function f was trained separately for each person and each target conversation session to avoid the severe drift caused by helmet shift. In Eq. 2, the head was assumed to rotate only along the vertical (y) axis (i.e. $\phi_x = \phi_z = 0$), because horizontal head rotations were frequent and large while vertical and in-plane rotations were infrequent and small, and thus hardly affected face localization.

The threshold for gazee recognition in the first-stage training was fixed at $\tau_d = 2.25°$, while the best threshold for the second stage, i.e. final gazee recognition, is assessed in Section V-A4. In RANSAC, 4,000 random sample sets were generated, and inlier judgment was performed with a threshold of 5°. Moreover, each sample set consisted of two samples as the minimum set needed to determine the mapping function Eq. 3. The original audio signals were normalized to the zero-mean-unit-variance in a pre-processing step, and τ_v was set at half of the mean power for each person.

D. Manually-generated data for performance evaluation

Although our system is almost fully automatic, we prepared three types of additional data manually for performance evaluation.

The first were samples for assessing the estimated gaze points. After the conversation sessions, each symmetric group member was asked to look at specified physical markers placed in space, as shown in Fig. 5. Five (horizontal) × three (vertical) markers were used in this study[8]. Consequently, in relation to the wearer's FOV, the markers were located roughly in the ±40° range for the x-axis and the ±23° range for the y-axis. The markers 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 from G^6, where the spatial range of the participants was the widest in the out-cams.

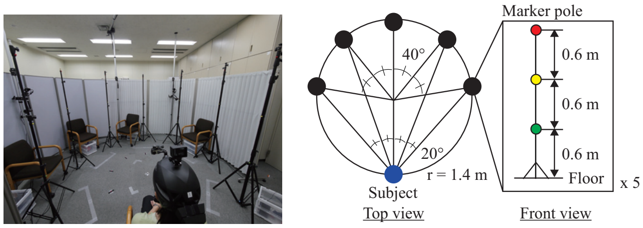


Fig. 5. Marker-based calibration scene for performance evaluation

Third, to evaluate the gazee and speaker recognition performance, the annotator also gave them frame-by-frame labels throughout the conversation sessions. As for the largest groups, that is G^6, to increase the reliability of the gazee annotation, two additional coders were employed and the final labels were determined by these three coders who used majority voting[9]. Moreover, the gazee annotation was accomplished by using videos recorded by additional cameras that were fixed in the environments; each of which captured the entire bodies of several participants.

E. Performance measures

The continuous estimations, namely gaze point estimation (and face and iris detection with the fundamental techniques), are evaluated in terms of the mean absolute errors (MAEs) in angle from the ground truth of the manually-provided marker coordinates (and face and iris image coordinates). The angle errors of the iris detection were calculated from pixel errors with the angle to pixel scale, which was obtained by using a rough estimate of the eyeball radius (r in Fig. 13) of 1.25 cm, with reference to [60], and that of the distance between the in-cam and the eyeball center (d in Fig. 13) of 17 cm. Gaze point estimation errors were obtained by using the physical markers, while the remaining performance characteristics were obtained by using the conversation data.

As performance measures for classification tasks, i.e. gazee recognition, mutual gaze recognition and speaker diarization, we use accuracy and precision/recall/F-score. Accuracy means the percentage of frames whose classification was the same as that obtained with human annotation, while F-score is the harmonic mean of precision and recall.

These scores were calculated as follows: A confusion matrix was first generated by using all the target data. See Table V as an example. Then, for a binary classification problem, i.e. mutual gaze recognition (mutual gaze or not), accuracy was calculated as (tp+tn)/(tp+tn+fp+fn), where tp, tn, fp and fn mean true positives, true negatives, false positives and false negatives, respectively. Precision and recall were obtained as tp/(tp+fp) and tp/(tp+fn), respectively. For multi-class problems, i.e. gazee recognition and speaker diarization, these scores were obtained as the average per-class scores [61]; that is, the score for each class or person

[8] Some additional markers were also placed, but they were omitted from this study; their horizontal angles (±60°) were too extreme for some subjects to look at them correctly, or the eyes were almost closed while looking these markers with our camera setting.

[9] The additional coders only judged frames where the first annotator gave a person's face label. For these three coders, Fleiss' kappa coefficient [58], which was used to show the mean pair-wise inter-coder agreement, was .88. This is judged as an excellent level according to [59].

TABLE II
LOCALIZATION ERRORS OF FACE CENTERS AND IRIS CENTERS

	MAE [degrees]		
	Hor	Ver	Mean
Face centers			
TLD [46] only	1.40	1.38	1.39
TLD [46] + Co-FPV (Eq. 2)	0.37	0.48	0.43
Iris centers			
Orientation-controlled FRST	2.28	3.37	2.83

number was first calculated individually, and then averaged. We used overall scores and per-group scores. They were calculated from the confusion matrices that were obtained by using all the data (called *overall* scores) and only the target group type, e.g. G^3, respectively. Moreover, only the fixation samples were used for the evaluation.

V. RESULTS

We first report the gaze point estimation errors, face localization errors, iris center localization errors, gazee recognition rates, and speaker identification rates. We then present generated gaze heat maps. We finally verify the rationale of the proposed framework.

A. Quantitative evaluation: accuracy assessment

We first evaluate the fundamental techniques, and then the proposed framework.

1) Fundamental techniques: Table II shows the localization errors of face centers and iris centers. The MAEs of the face center localization obtained with the TLD tracker were 1.4° for both axes The MAEs of iris center localization were 2.3° and 3.4° for both axes, respectively. We employed bias-removed errors to gain a better understanding of the performance, because we observed that the orientation-controlled FRST tends to bias the results compared with manual localization.

The overall precision and recall of our speaker detection were 0.82 and 0.43, respectively. The reason for the low recall is that we ignored low voices and overlapping speech among interlocutors. However, precision has a greater influence on the MAEs than recall in the first self-calibration stage. This is because precision roughly indicates the reliability of the training samples. Recall determines the required conversation length, but numerous frame-by-frame samples can be obtained in several minutes, as demonstrated later in Sections V-A3 and V-A4.

2) Face center localization in Co-FPV: Table II also shows that the proposed framework greatly improves coarse face center localization with the TLD tracker. The MAEs were reduced by 1.0° and 0.90° in the horizontal and vertical directions, respectively. Strictly speaking, we should note the ratio of the decrease on the horizontal axes (74%) to that on the vertical axes (65%), because Co-FPV refined the face center only along the horizontal axis in this study.

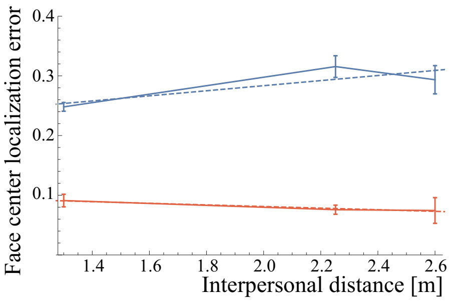


Fig. 6. Face center localization error (normalized by face-size) versus distance of target person from wearer: Red and blue indicate our Co-FPV and TLD only, respectively. Solid and dashed lines, respectively, indicate the actual data with its standard deviation and regression line.

For a more intuitive assessment, we further converted the absolute angle errors to face-size-normalized errors, because the target's face size in the out-cam images changes approximately linearly according to the distance between the target and the wearer in the real space, d'. The face-size-normalized errors were calculated as the pixel errors (angle errors multiplied by α_o) divided by the half target's face height estimated in the out-cam image[10]. The mean face-size-normalized error of our Co-FPV was 0.10, meaning that it localized face centers with the error of 10% of the face size.

Our method is robust against the distance to the target face. Figure 6 shows the relationship between the face-size-normalized error versus the distance of the target person from the wearer. The face-size-normalized error of our Co-FPV is regressed as $0.11 - 0.014d'$ (Pearson's correlation $r = -.98$). The slope is very small (one-ninth of the bias). For example, even when $d' = .46$ m, the shortest distance for good friends or family [62], the unnormalized error remains as small as 0.53° (24% increase from our settings, average distance of 1.9 m). The regression function of the TLD tracker was $0.20 + 0.043d'$, $r = 0.83$; both the bias and slope are much larger than those of our Co-FPV[11].

3) Gaze point estimation: Table III shows the MAEs of the gaze point estimation with three methods, namely self-calibration, which uses training samples where both the iris

[10] We estimated the actual face size as the half face size inferred by the TLD tracker, because we included non-face regions when initializing the tracker, as described in Section III-C2. Moreover, to focus on the basic property of our Co-FPV, here we only used samples that yielded errors of less than 3° when using the TLD tracker.

[11] The positive slope means that it yielded larger errors for distant people. This is probably because their resolution was very low (the captured faces were too small), compared with closer people.

TABLE III
LOCALIZATION ERRORS OF GAZE POINT IN DEGREES WITH COMPARISON WITH PREVIOUS METHODS

Method	Hor	Ver	Mean
Proposed method			
Self-calibration (Co-FPV)			
Two-step training	3.2	2.5	2.8
1st training stage only	3.5	2.5	3.0
1st training stage only w/ man. speaker	4.2	2.7	3.4
Manual-calibration			
Marker + man. iris	2.6	2.7	2.7
Marker + auto. iris	2.1	2.6	2.4
Marker + auto. iris (person-independent)	6.7	5.1	5.9
Passive-vision-based methods			
Self-calibration-based			
Chen and Ji [36]			1.8
Sugano et al. [37]			3.5
Alnajar et al. [40]			4.3
Person-independent			
Zhang et al. [63]			4.5
IR-lighting-based methods			
Manual-calibration-based			
Tsukada et al. [16]	0.7	0.9	0.8
Commercial products (from [36], [37])			1 - 2.7

Errors of previous methods are from the original papers.

TABLE IV
MEAN ABSOLUTE ERRORS OF GAZE POINT ESTIMATION IN DEGREES FOR DIFFERENT CONDITIONS

	Group size		
Conversation type	G^3	G^4	G^6
Self-introduction	3.0 (3.4/2.6)	2.8 (3.4/2.2)	3.0 (3.5/2.5)
Discussion	2.5 (2.5/2.5)	2.8 (3.2/2.3)	2.8 (3.2/2.4)

Values in bracket denote accuracies on x-axis (left) and y-axis (right).

positions and gaze points were automatically determined[12] (Co-FPV), a method using samples where the markers were manually determined but the irises were automatically localized ("Marker + auto. iris"), and a method using samples where both were manually localized ("Marker + man. iris")[13].

In Co-FPV, the second training stage succeeded in increasing the accuracy from an MAE of 3.0° in the first stage to an MAE of 2.8°. The improvement is mainly achieved on the x-axis. The difference in the x-axis of 0.3° is both statistically and practically significant, $t(57) = 3.2$, $p < .005$, $d = .22$ (Cohen's d), while the differences in both the overall (the mean of both axes) and y-axis MAEs are statistically significant $p < .05$ but practically trivial $d = .09$. Based on the average face width/height of around 3.0°, this suggests that it is possible to determine whether the wearer looked at face or body, although it is difficult to distinguish between different parts of the face, e.g. eyes from mouth.

"Marker + man. iris" and "Marker + auto. iris" yielded MAEs of 2.7°and 2.4°, respectively. Our iris localization outperforms manual localization. Although the MAE on the x-axis of Co-FPV is larger than both that of "Marker + man. iris" ($t(57) = 4.1$, $p < .001$, $d = .62$) and that of "Marker + auto. iris" ($t(57) = 7.1$, $p < .001$, $d = .90$), their MAEs on the y-axis are comparable ($t(57) = 1.1$, $p > .05$, $d = .21$, and $t(57) = 1.6$, $p > .05$, $d = .14$, respectively).

Furthermore, the mean increase rate in MAE of self-calibration that we obtained with manual calibration is also comparable to or smaller than that obtained with previous self-calibration methods. For example, the increase rate reported in [37] was 30% (from 2.7°to 3.5°), while that with our two-step training was 17% (from 2.4°to 2.8°). These similar trends support the validity of our evaluation. Moreover, these methods are difficult to compare directly due to the differences in the data and target situations, as mentioned in Section II-D. For example, most previous methods were evaluated in a narrower range, especially on the x-axis, than our evaluation range, i.e. ±40°.

To assess the interpersonal difference in the parameters of the mapping function, "Marker + auto. iris (person-independent)" in Table III represents the errors when the parameters were determined by a leave-one-subject-out cross-validation scheme; i.e. the parameters were set at the means of those obtained in "Marker + auto iris" for other participants with the same camera resolution. The large errors suggest the need for person-specific calibration.

Table IV shows the effect of conversation type and group size on gaze point estimation. Our method is robust against both conversation type and group size. A two-way ANOVA revealed that the main effects of both factors were neither statistically ($p > .05$) nor practically significant ($\eta^2 < .01$), according to Cohen's criteria [64]. A reason for the high performance with the discussion session is that more widely distributed training samples were available due to more frequent turn-changing than in the self-introduction sessions.

Another possible reason is the conversation length. Thus, to validate its effect, Figures 7 and 8 show how the errors vary when only a part of the conversation data taken from the beginning and the end, respectively, is used for training. In summary, 5-6 min was sufficient for our two-stage training; the errors reached the lower bound for all conversations on both axes at 5-6 min in these figures. If there is a strong demand for a shorter conversation but a larger error is acceptable, we have the option of using only the first training stage, which requires only 3-5 min.

Figures 7 and 8 suggest that the proposed method can handle all types of conversations, namely ordered and unordered conversations (defined in Section IV-A) and a mixture of the two. First, Fig. 7 suggests that our method

[12] We observed that the aligned eye position was slightly biased between the conversation sessions and marker-based calibration due to differences in the spatial distribution of the gaze direction. Because one of our main proposals is sample selection rather than accurate helmet slip compensation, we removed the bias using the following steps: first, the bias was calculated for a marker point, and then removed from all the remaining marker points. This yields (the number of markers - 1) bias estimates for each point. The mean of these estimates is considered to be the bias at that point. Note that the bias was not directly calculated from the target point, and the slope factor estimates were not affected by this process at all. The bias removal was just applied to this evaluation, not applied to gazee/mutual gaze recognition in Section V-A4. The validity of this evaluation is supported by the training failure with narrow windows in shown Fig. 7 and the high classification performance levels in Section V-A4. To alleviate the helmet shift problem, the idea of sliding-window-based adaptive calibration, as described in [38], would be beneficial.

[13] For a qualitative assessment, a movie is available from http://www.kecl.ntt.co.jp/people/kumano.shiro/research/gazeanalysis.htm.

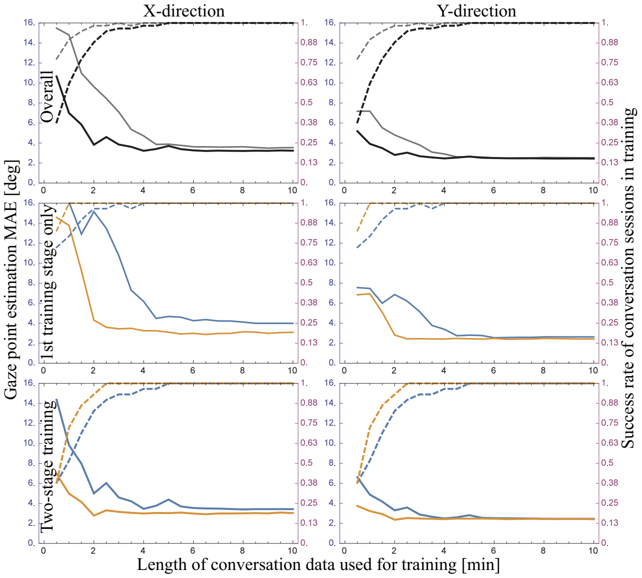


Fig. 7. Change of gaze point estimation MAE in x-direction (left column) and y-direction (right column) when varying the length of the conversation data used for training (horizontal axes); (Top) overall, (middle) first training stage only, and (bottom) two-stage training: Solid lines denote the MAE (scaled by the left vertical axis), while dotted lines indicate the success rate of conversation sessions in training (scaled by the right vertical axis); i.e. if the success rate is not one, it means that some sessions failed in the training and they were not considered in calculating the mean MAE. Thickness represents training stages: first stage (thin), and two stages (thick). Color indicates conversation types: overall (black), self-introduction (blue), or discussion (orange).

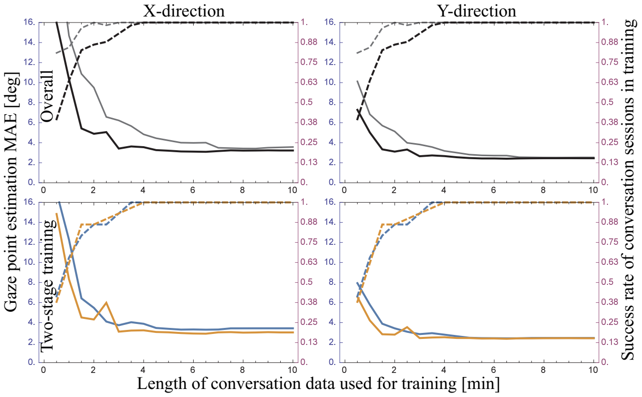


Fig. 8. Change of overall gaze point estimation MAE when varying the length of the conversation data used for training; (Top) overall, and (bottom) two-stage training: The only difference from Fig. 7 is that the end of the window was fixed to the end of the conversations, while, in Fig. 7, the start of the window was fixed to the beginning of the conversations. The middle panels of Fig. 7 are not included here.

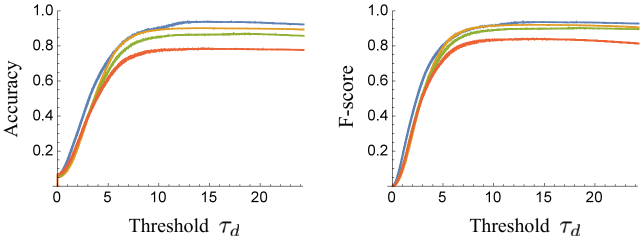


Fig. 9. Accuracy (left) and F-score (right) of gazee recognition versus τ_d: Blue, Green, Orange and Red indicate G^3, G^4, $^*G^4$, and G^6, respectively. Best accuracies are .94, .86, .90, and .78, respectively, at $\tau_d = 15.0°$. Best F-scores are .94, .90, .92, and .84, respectively, at $\tau_d = 14.4°$.

TABLE V
CONFUSION MATRIX OF GAZEE RECOGNITION AT THE BEST F-SCORE

	1	2	3	4	5	6	0
1	84952	111	78	0	0	182	10262
2	566	66001	234	16	0	13	8430
3	0	237	78692	66	0	0	8760
4	7	0	171	39732	54	4	6327
5	0	18	0	268	16302	78	3640
6	106	9	0	5	215	13663	3124
0	1505	1532	2126	1381	788	814	14155

Rows and columns indicate human annotation and our method's estimation. Classes in the rows and columns (0 to 6) denote person numbers, which are shared among groups (i.e. "1" means the first person in all groups). "0" represents not-person.

can handle ordered conversations (see the blue lines) and mixed conversations (see the orange lines). Moreover, it is natural with respect to the turn-taking structure that the self-introduction session required a longer conversation (the fact that the blue lines are almost all above the orange lines in Fig. 7). This is because everyone's speech lasted about 1.3 min and thus it took a long time to reach the last speaker's turn, while it lasted 34 sec at the beginning of the discussion sessions, as described in Section IV-A. Second, similar results were obtained in Fig. 8 even by training the model using only the *tails* of the conversations, i.e. the *end* of the window was fixed at the *end* of the conversation. Although we expected unordered conversation to prove a more challenging scenario, this suggests that our method can also handle unordered conversations.

4) Gaze/mutual-gaze recognition: Figure 9 shows the accuracy and F-score curves of the gazee recognition for each group size when τ_d is gradually changed. The best overall accuracy of .86 and the best overall F-score of .89 (precision = .91 and recall = .87) were obtained at $\tau_d = 15.0°$ and $\tau_d = 14.4°$, respectively[14]. It is natural that the performance degrades as the group becomes larger, because the task is a N-class problem.

These results suggest the upper limit of human coding, although they can be changed slightly if more coders are employed. First, the best threshold angles are much larger than the errors in the gaze point estimation shown in Section V-A3. These angles are in fact similar to those reported in [12], which reported that the human accuracy when distinguishing targets separated by a visual angle of 8°-10° is around 40%. Second, the recognition performance strongly depends on the group size, although the gaze point estimation does not.

The best thresholds obtained separately for each group type reveal the characteristics of human annotation. The F-score results were 14.9°, 18.6°, 14.5°, and 14.4°for G^3, G^4, $^*G^4$, and G^6, respectively. These thresholds are, except for G^3 (the most sparse arrangement), slightly smaller than a

[14] We also have tried an ellipse-based thresholding, where the thresholds are different for x- and y-axes, and obtained the best thresholds similar to that of the circular thresholding.

TABLE VI
ACCURACY OF GAZEE RECOGNITION

	Group size					
Method	2	G^3	G^4	$^*G^4$	5	G^6
Co-FPV	-	.94	.86	.90	-	.78
Fixed-cam-based methods						
Gorga+ [21]	-	-	-	.82	-	-
Stiefelhagen+ [20], [65]	-	-	.76	-	.69	-
Ba+ [29]	-	-	.56	-	-	-
Mora+ [18]	.86	-	-	-	-	-

Dyadic and five-person groups are included to stress the strong relationship between the accuracy and group size; i.e. poorer performance in larger groups. The accuracy for [29] was obtained as a seven-class classification task, including other objects, such as a screen. To obtain a fair comparison, we estimated the performance of our approach for a similar task, that is the performance of our approach for a seven-party conversation, by fitting a linear model. The estimated accuracy was .73, which is much higher than that reported in [29].

TABLE VII
F-SCORE OF MUTUAL GAZE RECOGNITION

	Group size				
Method	2	G^3	G^4	$^*G^4$	G^6
Co-FPV (self-calibration)	-	.94	.85	.89	.68
Manual-calibration-based methods					
Ye+ [66]	.76	-	-	-	-
Martinez+ [19]	-	.87	-	-	-

half of the (minimum, for the asymmetric groups) interval angles between participants from the viewpoint of each wearer. The half intervals are 30°, 22.5°, 15°, and 15° for these arrangements. Furthermore, Table V shows the confusion matrix at $\tau_d = 14.4°$. 95% of the confusions are those between person and not-person, and 54% of them were created in G^6, the most crowded setting.

Table VI compares the mean accuracies with previous studies' reports. Our results exceed previous ones for all group sizes, although these previous methods used human annotation for model training [18], [29], marker-based calibration [21], and additional sensors [20]. These rates are difficult to compare statistically due to differences in the experimental settings. Note that these measures are also sensitive to class skewness, i.e. the ratio of positive and negative samples (in our case, the frequency with which interlocutors look at others).

Table VII shows the F-score for mutual gaze recognition. The overall F-score is .87. Although the performance decreased slightly from that of individual gazee recognition, it is still higher than that in previous manual-calibration-based studies [19], [66]. The sudden drop at G^6 again suggests the limitations of human annotation for medium-sized parties. Finally, Table VIII shows the accuracies; the overall accuracy is very high at .98. However, the fact that the accuracy is much higher than the F-score is not very informative, because the mutual gaze samples are highly skewed and the accuracy uses true negative samples, i.e. those correctly recognized as no mutual gazes, which were 92% (79%, 92%, 87%, and 97% for G^3, G^4, $^*G^4$, G^6, respectively) of our data; while the F-score does not take account of them. For more details regarding the effect of class imbalance on the performance measures, see [61].

TABLE VIII
ACCURACY OF MUTUAL GAZE RECOGNITION

	Group size			
Method	G^3	G^4	$^*G^4$	G^6
Co-FPV (self-calibration)	.97	.97	.98	.98

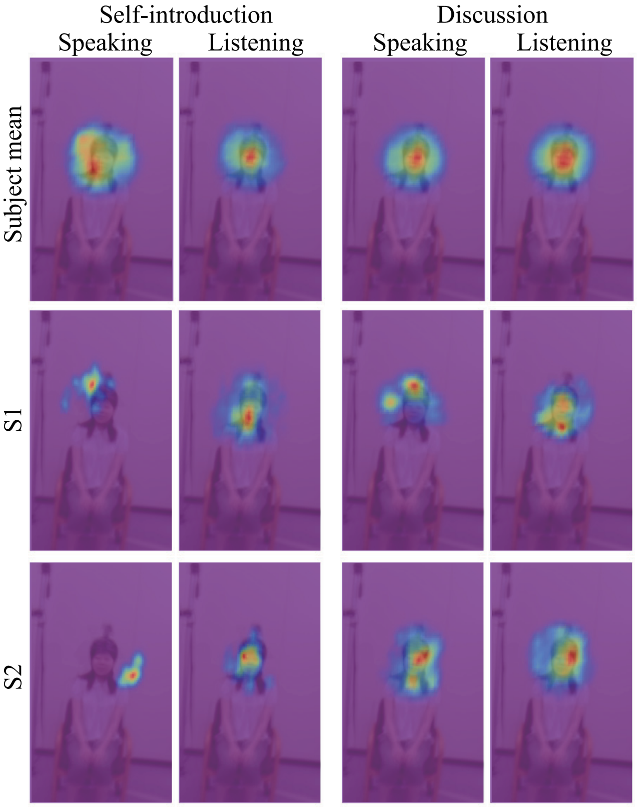


Fig. 10. Gazee-centered heat maps: Top) the mean map of all participants, middle and bottom) the maps of two subjects. Although the gaze points are mainly on the face, S1 while speaking for both conversation sessions and S2 while speaking for the self-introduction session avoided looking at other's face. The background image is just an illustration; it means the image of a participant captured from the out-cam of another person who sat in front of the participant, while the heat maps were generated from the gaze toward $N - 1$ persons.

B. Qualitative assessment: gaze heat maps

Figure 10 shows the mean gazee-centered heat maps of all the participants and two distinctive persons (denoted S1 and S2). They are separated into those obtained under speaking and listening conditions, and from the self-introduction and discussion sessions. It is natural that the gaze point while listening is mainly centered on the face, i.e. the red area, meaning the most frequent point is on the face, because we imposed such a constraint during training.

The gaze points while speaking thus provide notable differences between the participants that suggest the effect of conversation type, social pressure, and participants' personality and emotional state, as previously suggested in [67]. For example, the more distributed gaze points for the self-introduction than for the discussion session while speaking suggest higher pressure on the conversation. Focusing of the interpersonal differences, S1 avoided looking at others'

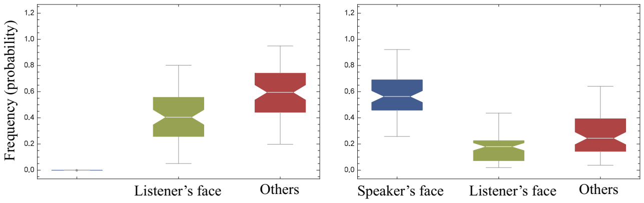


Fig. 11. Frequencies of gaze targets of speaker (left) and listener (right). The most salient gazee is the speaker.

faces for both conversation sessions; i.e. she appeared to avoid eye contact with the listeners. S2 showed a similar tendency only in the self-introduction session. Although further psychological analysis is required, these results are promising as regarding demonstrating the effectiveness of the proposed method for such studies.

C. Verification of basis of proposed framework

We verify the basis of the proposed framework, including the reason for it working well. We first test our basic assumption, i.e. listeners are likely to look at the speaker's face, and then investigate how our formulation, Eq. 1 and Eq. 3, work under this conversation rule.

1) Validity of our basic assumption: Figure 11 shows the mean frequencies at which speakers and listeners were looked at obtained using all our conversation data. The frequencies were created by using the human annotation. It demonstrates that speakers showed clearly dominant gaze targets. We test if their differences are statistically/practically significant.

As regards listeners' gaze targets, we use a three-way ANOVA; the three factors are gaze target (speaker's face, (an)other listener's face, or others), group size N, and conversation type (self-introduction or discussion). Each sample has the probability (normalized frequency) of the occurrence of one person, and there are 66 samples. It reveals that the only main effect is gaze target, $p < .001$, $\eta^2 = .63$ (large effect). The other effects and all the interactions are non-significant ($p \geq .05$) or have only small or trivial effects ($\eta^2 < .06$). Post-hoc paired-t tests with Bonferroni correction reveal that listeners preferred the speaker's face to (an)other listener's faces, $t(65) = 16.5$, $p < .001$, $d = 3.3$ (large effect), and to other targets, $t(65) = 8.8$, $p < .001$, $d = 2.1$ (large effect), respectively. Moreover, (an)other listener's face was looked at less frequently than other targets, $t(65) = 4.2$, $p < .001$, $d = 0.8$ (large effect). These results suggest that our basic assumption is true, and they match those found in previous studies [13], [15]. Our new findings are that the listener's gaze tendency is not strongly dependent on either group size or conversation type.

The trend regarding the speakers' gaze was very similar to that of the listeners'. A three-way ANOVA reveals that the only main effect is gaze target, $p < .001$, $\eta^2 = .20$ (large effect). The other effects and all the interactions are non-significant or have only small or trivial effects. Post-hoc paired-t tests reveal that the listener's face was looked at less frequently than other targets, $t(65) = 4.1$, $p < .001$, $d = 1.0$ (large effect). Furthermore, the listener's face is not useful for self-calibration even while the target wearer is speaking. It was not clear who the gazee was among the listeners, though the dominance effect, i.e. the effect whereby a more dominant person is looked at more frequently than a less dominant person [68], does exist.

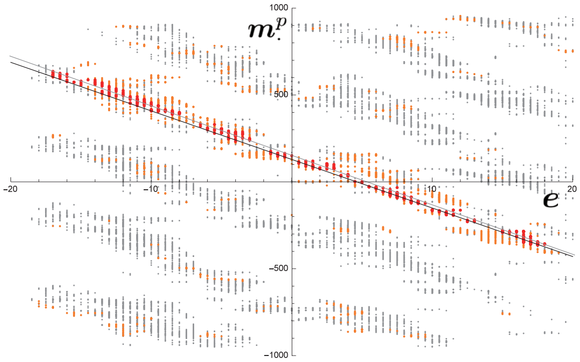


Fig. 12. Plot of horizontal position of iris in the in-cam images (x-axis) and speaker face in the out-cam images (y-axis) for a participant (p): Orange, red and gray dots denote speaker faces identified by the audio signals, speaker faces identified at the second training stage, and all others' faces, respectively. Gray and black lines represent the fitted mapping function at the first and second training stages, respectively.

2) Spatial characteristics of our data: Figure 12 shows the relationship between the horizontal face positions in the out-cams on the y-axis and the corresponding wearer's iris centers in the in-cams on the x-axis. Note that the speaker faces (orange dots) mostly lie on a single line, although other faces (gray dots) form several clusters. Equation 3, drawn as a gray line (the first stage) and a black line (the second stage), well approximates this mapping in conjunction with the outlier removal in solving Eq. 1.

This figure also suggests that simply fitting the line for the cluster closest to the origin with or without distinguishing the speaker would be sufficient in this case. However, it is not applicable to every camera configuration and participant arrangement[15]. The clear separation between the gray dots might be diminished in other scenarios, e.g. in-motion scenarios, where everyone moves freely. In contrast, the speaker's line, formed by orange dots, would remain, unless the wearer frequently looked at a listener whose horizontal position is close to the speaker's. The inliers were further cleaned at the second training stage (red dots).

VI. Discussion

The experiment demonstrated the basic validity of Co-FPV. However, several issues remain.

A. Usability

In its current form, our setting using a helmet with two cameras does not appear very usable in the wild, unlike

[15] We have already tried to exploit this prior knowledge in [41]. However, we later found that RANSAC works well in our settings without such prior knowledge.

other monocular-camera-based approaches, e.g. [22], [63]. However, the dual-view system is useful at least in the laboratory in the sense that it also has the potential to measure other nonverbal behaviors, such as facial expressions and head gestures, which are other major focuses of affective computing and behavioral psychology. If usability is the crucial factor in terms of the research/application, glasses-type camera devices, such as that reported in [16], would be the first choice. Non-camera devices, e.g. electrooculography [69], would be also potential candidates. Even with these devices, the proposed framework could be used to jointly automate their calibration steps, although evaluation with an appropriate mapping function is needed. On the other hand, if social cognition is the main focus, such helmet-type devices are more beneficial. This is because the frame of the glasses would change another's impression regarding the appearance of the wearer's face and/or facial expressions [70].

B. Applicability

This study made several assumptions. However, only the two are crucial to our framework: people often converse with each other, and they are likely to look at the face of the speaker. These assumptions would be largely true for a variety of conversation scenarios. However, in some more challenging cases, the latter assumption would not hold true, and so our method would not work well in its current form. One such case is where the target is a person with autism who tends to gaze at the other's body [33]. Another case is where there are salient objects in the environment, e.g. a monitor displaying presentation slides. Moreover, our framework has the potential to handle meeting with paper documents by eliminating occasions where the participant looks downwards from the training samples.

Some of the remaining constraints, e.g. standing or in-motion conversations tackled in [22], [23], could be technically relaxed by introducing crowd tracking techniques, especially tracking-by-detection with data association, e.g. [71]; and manual camera (internal parameter) calibration. Moreover, from a psychological viewpoint, such a free dynamic scenario makes it difficult to control the experiment, and it remains unclear whether our key assumptions remain valid even in such a scenario.

Furthermore, we discuss the applicability when glasses are worn, because no one wore glasses in this study. Both the current iris and face localization methods are probably robust against glasses to a certain extent. As for iris detection, the effect of the glasses' frames would be alleviated by limiting the votes of pixels in the FRST by their curvatures. Specular reflection on glasses is unlikely to occur when shooting from above. The effect of glasses on face localization is expected to be less severe, because both the TLD tracker and our Co-FPV are holistic methods, i.e. they use the entire face region.

C. Gaze model

This paper dealt with gaze point as a two-dimensional point in an 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 [72], would increase the accuracy. This would visualize/describe three-dimensional gaze interactions among people.

D. 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 [73]. Additionally, other pair-wise metrics, e.g. gaze following, or group-wise metrics, would be applicable in social interactions. Determining the best gaze metrics, in conjunction with an analysis of pupil size, as an indicator of cognitive load [74] or other affective states [75], is another issue but it is beyond the scope of this paper.

VII. Conclusions and Future Directions

We introduced the Co-FPV framework for automatic conversation analysis, where captured audio-visual signals are gathered in a centralized system, and the fundamental components required for group gaze analysis are jointly and effectively processed. Each participant's gaze tracker is self-trained off-line by automatically selecting training samples based on the conversational rule. This estimation approach yields a gazee-centered heat map for each interlocutor. An exhaustive experiment using three to six-member 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 used to obtain emotion data. Our assumption is that first-person view images make it easier for the subject (or external observers) to accurately recall (or read) his/her emotion felt at that time, because the first-person perspective (or perspective taking for the observers) plays an important role in these processes [76], [77]. The evaluation of this assumption is a future task.

Appendix

The mapping function in Eq. 3 is obtained as follows.

First, we only consider the following parameters in Fig. 13, which shows the geometrical relationship between the cameras and the eyeballs: 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 are omitted by assuming that the locations of the centroid of the eyeballs and the focal point of the out-cam are the same, both the in- and out-cams are horizontally aligned (i.e. no roll rotations), and the yaw and pitch angle of the out-cam relative to the eyeball are zero. These assumptions are guaranteed 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

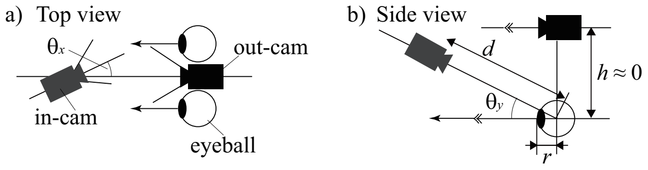


Fig. 13. Assumed geometrical relationship of cameras and eye

the person in front of the wearer is located at the image center.

When we consider the x-axis in Fig. 13, the 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 a weak-perspective projection. We assume that $d \gg r$ 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

$$g_x = \alpha_o[\arcsin\{d/r \cdot \tan(-e_x/\alpha_i)\} + \theta_x]. \quad (5)$$

The derivation for the y-axis is the same except for the signs of e_x and e_y. Thus, the mapping function f forms as:

$$f(\boldsymbol{e}) = \alpha_\mathrm{o}[\arcsin\{d/r \cdot \tan((-e_x, e_y)^\mathrm{T}/\alpha_\mathrm{i})\} + (\theta_x, \theta_y)^\mathrm{T}], (6)$$

where α_i is in-cam's scale factor that relates degrees to pixels in both axes. The sign of e_x is changed because the x-axis is flipped in the in-cam, as shown in Fig. 13.

Finally, a first-order Taylor series approximation of the right hand side of Eq. 6 around $(0, 0)$ makes the mapping function f a similarity transformation

$$\boldsymbol{g} = f(\boldsymbol{e}) \approx \bar{f}(\boldsymbol{e}) = \begin{pmatrix} -a & 0 & b_x \\ 0 & a & b_y \end{pmatrix} \boldsymbol{e}'. \quad (7)$$

This is equivalent to Eq. 3.

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] S. Okada, O. Aran, and D. Gatica-Perez, "Personality trait classification via co-occurrent multiparty multimodal event discovery," in *Proc. ICMI*, 2015.

[4] W. Dong, B. Lepri, F. Pianesi, and A. Pentland, "Modeling functional roles dynamics in small group interactions," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 83–95, 2013.

[5] L. S. Nguyen, D. Frauendorfer, M. S. Mast, and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1018–1031, 2014.

[6] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1075–1089, 2014.

[7] S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato, "Analyzing interpersonal empathy via collective impressions," *IEEE Trans. Affective Computing*, 2015.

[8] A. Marcos-Ramiro, D. Pizarro, M. M. Romera, and D. Gatica-Perez, "Let your body speak: Communicative cue extraction on natural interaction using RGBD data," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1721–1732, 2015.

[9] V. Peruffo Minotto, C. Rosito Jung, and B. Lee, "Multimodal multichannel on-line speaker diarization using sensor fusion through SVM," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1694–1705, 2015.

[10] R. Valenti and T. Gevers, "Accurate eye center location through invariant isocentric patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1785–1798, 2012.

[11] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[12] T. Gao, D. Harari, J. Tenenbaum, and S. Ullman, "When computer vision gazes at cognition," in *Tech. Rep. Center for Brains, Minds, & Machines*, 2014.

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

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

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

[16] A. Tsukada, M. Shino, M. S. Devyver, and T. Kanade, "Illumination-free gaze estimation method for first-person vision wearable device," in *Proc. IEEE ICCV Workshops*, 2011.

[17] S. Nihonyanagi, Y. Hayashi, and Y. I. Nakano, "Analyzing co-occurrence patterns of nonverbal behaviors in collaborative learning," in *Proc. Workshop GazeIn*, 2014, pp. 33–37.

[18] K. A. Funes Mora and J.-M. Odobez, "Person independent 3D gaze estimation from remote RGB-D cameras," in *Proc. IEEE ICIP*, 2013, pp. 2787–2791.

[19] F. Martinez, A. Carbone, and E. Pissaloux, "Combining first-person and third-person gaze for attention recognition," in *Proc. IEEE Int'l Conf. FG*, 2013, pp. 1–6.

[20] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 928–938, 2002.

[21] S. Gorga and K. Otsuka, "Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection," in *Proc. ICMI-MLMI*, 2010.

[22] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir, "Automatic editing of footage from multiple social cameras," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 81:1–11, 2014.

[23] 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.

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

[25] 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.

[26] R. Yonetani, K. M. Kitani, and Y. Sato, "Recognizing micro-actions and reactions from paired egocentric videos," in *Proc. IEEE CVPR*, 2016.

[27] F. Chen, D. Delannay, and C. D. Vleeschouwer, "An autonomous framework to produce and distribute personalized team-sport video summaries: A basketball case study," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1381–1394, 2011.

[28] 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.

[29] S. O. Ba and J.-M. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 101–116, 2011.

[30] A. Betancourt, P. Morerio, C. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 744–760, 2015.

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

[32] H. S. Park and J. Shi, "Social saliency prediction," in *Proc. IEEE CVPR*, 2015.

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

[34] 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.

[35] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, 2010.

[36] J. Chen and Q. Ji, "Probabilistic gaze estimation without active personal calibration," in *Proc. IEEE CVPR*, 2011, pp. 609–616.

[37] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 329–341, 2013.

[38] D. Perra, R. Kumar Gupta, and J.-M. Frahm, "Adaptive eye-camera calibration for head-worn devices," in *Proc. IEEE CVPR*, 2015.

[39] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *Proc. IEEE ICCV*, 2013, pp. 3216–3223.

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

[41] S. Kumano, K. Otsuka, R. Ishii, and J. Yamato, "Automatic gaze analysis in multiparty conversations based on collective first-person vision," in *Proc. Int'l Workshop EmoSPACE*, 2015.

[42] G. Loy and A. Zelinsky, "A fast radial symmetry transform for detecting points of interest," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 959–973, 2003.

[43] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

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

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

[46] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, 2012.

[47] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, 2009.

[48] K. Miyamoto, "Fish eye lens," *J. Opt. Soc. Am.*, vol. 54, no. 8, pp. 1060–1061, 1964.

[49] C. Hughes, P. Denny, M. Glavin, and E. Jones, "Equidistant fish-eye calibration and rectification by vanishing point extraction." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2289–2296, 2010.

[50] R. I. Hartley, "In defense of the eight-point algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 6, pp. 580–593, 1997.

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

[52] Y. Wang and B. Yuan, "A novel approach for human face detection from color images under complex background." *Pattern Recogn.*, vol. 34, pp. 1983–1992, 2001.

[53] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recogn.*, vol. 40, no. 3, pp. 1106–1122, 2007.

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

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

[56] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via l0 gradient minimization," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 174:1–174:12, 2011.

[57] J. Panero and M. Zelnik, *Human Dimension and Interior Space: A Source Book of Design Reference Standards*. New York: Watson-Guptill, 1979.

[58] J. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

[59] J. L. Fleiss, *Statistical Methods for Rates and Proportions*, 2nd ed. John Wiley, New York, 1981.

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

[61] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, 2009.

[62] E. T. Hall, *The hidden dimension*. New York, US: Doubleday & Co, 1966.

[63] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE CVPR*, 2015.

[64] J. Cohen, "A power primer," *Psychol. Bull.*, vol. 112, no. 1, pp. 155–159, 1992.

[65] R. Stiefelhagen, "Tracking and modeling focus of attention in meetings," Ph.D. dissertation, Karlsruhe Institute of Technology, 2002.

[66] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg, "Detecting eye contact using wearable eye-tracking glasses," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 699–704.

[67] 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.

[68] J. F. Dovidio and S. L. Ellyson, "Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening," *Social Psychology Quarterly*, vol. 45, pp. 106–113, 1982.

[69] Y. Chen and W. S. Newman, "A human-robot interface based on electrooculography," in *Proc. IEEE Int'l Conf. Robotics and Automation*, 2004, pp. 243–248.

[70] Åke Hellström and J. Tekle, "Person perceptions through facial photographs: Effects of glasses, hair, and beard on judgments of occupation and personal qualities," *European Journal of Social Psychology*, vol. 24, no. 6, pp. 693–705, 1994.

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

[72] H. S. Park, E. Jain, and Y. Sheikh, "3d social saliency from head-mounted cameras," in *Proc. NIPS*, 2012, pp. 431–439.

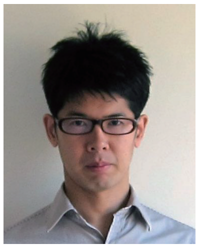
[73] A. Poole and L. J. Ball, *Encyclopedia of Human Computer Interaction*. Pennsylvania: Idea Group, 2004, ch. Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects.

[74] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," *Psychol. Bull.*, vol. 91, no. 2, pp. 276–292, 1982.

[75] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *Int. J. Hum.-Comput. Stud.*, vol. 59, no. 1-2, pp. 185–198, 2003.

[76] 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.

[77] 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.



Shiro Kumano received a PhD degree in Information Science and Technology from the University of Tokyo in 2009. He is currently a research scientist at NTT Communication Science Laboratories, and an honorary research associate at University College London. His research interests include computer vision, and affective computing, especially in relation to facial expression recognition and the automatic understanding of empathy. He received the ACCV 2007 Honorable Mention Award, the ICMI 2014 Outstanding Paper Award etc. He has served as an organizing committee member of the IAPR International Conference on Machine Vision Applications. He is a member of the IEEE, and IEICE.

Kazuhiro Otsuka received his B.E. and M.E. degrees in electrical and computer engineering from Yokohama National University in 1993 and 1995, respectively. He joined the NTT Human Interface Laboratories, Nippon Telegraph and Telephone Corporation in 1995. He received his Ph.D. in information science from Nagoya University in 2007. He was a distinguished invited researcher at Idiap Research Institute in 2010. He is now a senior research scientist/supervisor in the NTT Communication Science Laboratories. His current research interests include communication science, multimodal interactions, and computer vision. He was awarded the IAPR Int. Conf. on Image Analysis and Processing Best Paper Award in 1999, the Outstanding Paper Awards of ACM Int. Conf. on Multimodal Interfaces (Interaction) in 2007, 2012, and 2014, the Meeting on Recognition and Understanding (MIRU) 2009 Excellent Paper Award, the IEICE Best Paper Award 2010, the IEICE KIYASU-Zen'iti Award 2010, and others. He is a member of the IEEE, the IEICE and the IPSJ.

Ryo Ishii Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories. He received the B.S. and M.S. in Computer and Information Sciences from Tokyo University of Agriculture and Technology in 2006 and 2008, respectively, and the Ph.D. degree in informatics from Kyoto University in 2013. He joined NTT Cyber Space Laboratories in 2008. He moved to NTT Communication Science Laboratories in 2012 and NTT Media Intelligence Laboratories in 2016. He was also an invited researcher at Seikei University from 2011 to 2013. His current research interests include communication science, multimodal interactions, and human-computer interaction. He received the FY 2014 IEICE HCG Research Award and the ACM Int. Conf. on Multimodal Interaction 2014 Outstanding Paper Award. He is a member of IEICE and JSAI.

Junji Yamato received the B.E., M.E., and Ph.D. degrees from the University of Tokyo in 1988, 1990, and 2001, respectively, and the S.M. degree in electrical engineering and computer science from Massachusetts Institute of Technology in 1998. He is a professor of the Department of Information systems and applied mathematics, Faculty of Informatics, Kogakuin University in Tokyo. His research interests include computer vision, pattern recognition, human-robot interaction, and multiparty conversation analysis. He is a senior member of IEEE, IEICE, and a member of the Association for Computing Machinery.