Recognizing Micro-Actions and Reactions from Paired Egocentric Videos

Ryo Yonetani
The University of Tokyo
Tokyo, Japan
yonetani@iis.u-tokyo.ac.jp

Kris M. Kitani
Carnegie Mellon University
Pittsburgh, PA, USA
kkitani@cs.cmu.edu

Yoichi Sato
The University of Tokyo
Tokyo, Japan
ysato@iis.u-tokyo.ac.jp

Abstract

We aim to understand the dynamics of social interactions between two people by recognizing their actions and reactions using a head-mounted camera. Our work will impact several first-person vision tasks that need the detailed understanding of social interactions, such as automatic video summarization of group events and assistive systems. To recognize micro-level actions and reactions, such as slight shifts in attention, subtle nodding, or small hand actions, where only subtle body motion is apparent, we propose to use paired egocentric videos recorded by two interacting people. We show that the first-person and second-person points-of-view features of two people, enabled by paired egocentric videos, are complementary and essential for reliably recognizing micro-actions and reactions. We also build a new dataset of dyadic (two-persons) interactions that comprises more than 1000 pairs of egocentric videos to enable systematic evaluations on the task of micro-action and reaction recognition.

1. Introduction

The dynamics of social interactions between two people can be decomposed into a sequence of action and reaction pairs (such as pointing and sharing a point of attention, gesturing and nodding in agreement, or laughing and gesturing disagreement) to convey to each other a sense of their internal states. Our everyday interactions even include *micro-actions* and *micro-reactions* in which only subtle body motion is apparent, such as slight changes in focus of attention (small movement of the head in response to pointing), subtle nodding, or small hand actions. The ability to understand interaction dynamics with such micro-behaviors is important for human-to-human communications, as this mode of non-verbal communication is perhaps our primary means of understanding and expressing our internal state. Towards understanding the deeper complexities of social interaction dynamics, this work attempts to take the first step by developing a method to recognize micro-actions and reactions.

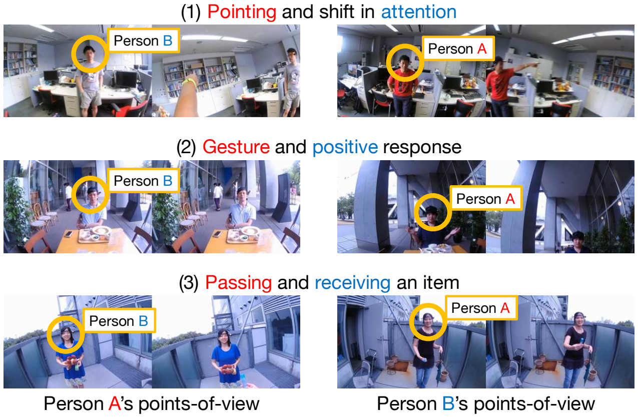

(1) Pointing and shift in attention
Person B
Person A
(2) Gesture and positive response
Person B
Person A
(3) Passing and receiving an item
Person B
Person A
Person A's points-of-view
Person B's points-of-view

Figure 1. Challenges of recognizing micro-actions. Slight head motion of person B induces only slight local motion in the person A's points-of-view in (1) and (2). Hand motion by person A is difficult to observe from the A's points-of-view in (2) and (3).

To enable such recognition ability, we show in this work that it is critical to have access to a pair of egocentric videos taken by two interacting parties. Particularly, we focus exclusively on dyadic (*i.e.*, two-person) interactions and assume that both people are equipped with a head-mounted camera. In this setting, each person always has a *first-person* point-of-view (POV) observation of one's self in one's own video and a *second-person* POV observation of the self in another video. For example, Figure 1(1) shows person A pointing from both his own POV (left) and the POV of person B (right). In this way, egocentric videos are advantageous from a sensing perspective since the head motion and hand motion of camera wearers are often observed clearly in such videos, making it possible to perform various forms of first-person action recognition [7, 9, 13, 18, 25, 27, 30]. They can also be used to see the behavior of other people up-close from the second-person POV [1, 2, 3, 8, 29, 44, 45].

One key observation to use a pair of egocentric videos is that a head-worn camera naturally amplifies subtle head motion and hand motion needed to recognize micro-actions

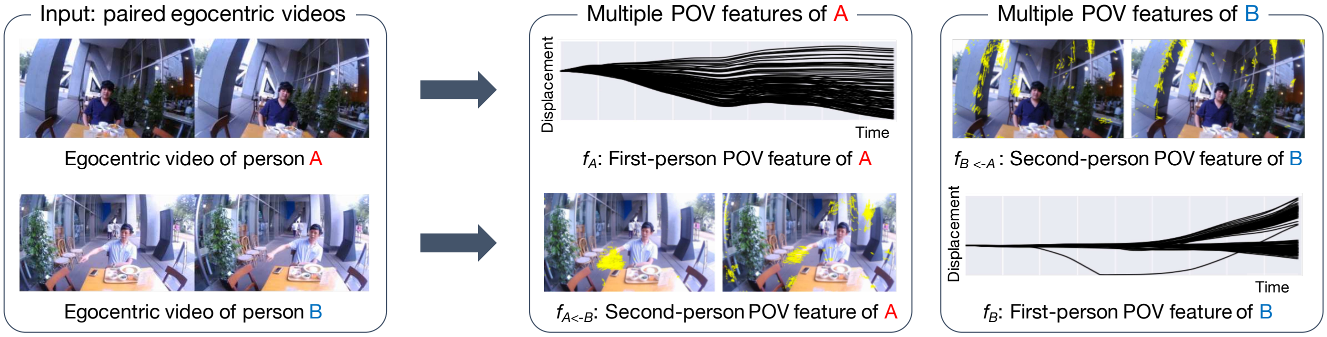

Input: paired egocentric videos
Egocentric video of person A
Egocentric video of person B
Multiple POV features of A
Displacement
Time
f_A: First-person POV feature of A
$f_{A<-B}$: Second-person POV feature of A
Multiple POV features of B
$f_{B<-A}$: Second-person POV feature of B
Displacement
Time
f_B: First-person POV feature of B

Figure 2. Our approach. Paired egocentric videos recorded by persons A and B are used to provide first-person and second-person POV features of both A and B, which are complementary and essential for recognizing micro-actions and reactions. Cumulative displacement patterns [27] and improved dense trajectories [39] are respectively visualized as examples of the first-person and second-person features.

and micro-reactions. For example, slight changes in focus of attention or subtle nodding cannot be adequately recognized from a second-person POV because they only induce slight variations in local motion (*e.g.*, person B seen in A's POV videos in the left of Figure 1(1)(2)). However, if we can gain access to the first-person POV of B, a small change in head pose translates to a large change in optical flows (the right of (1)(2)), making it possible to detect such micro-reactions. By contrast in Figure 1(2)(3), while hand motion of person A is not always large enough to be observed in the first-person POV (the left of (2)(3)), it is often more visible in the second-person POV including that person up-close (the right of (2)(3)).

Another key observation that motivates our work is that micro-actions and reactions are often correlated and best recognized when one has access to egocentric videos of both interacting parties. For example, in Figure 1(1), person A performs the action of *pointing*, which induces a micro-reaction of a shift in *attention* by person B. In other words, the context of *pointing* allows us to expect a responsive change in *attention*. Figure 1(2)(3) show other action-reaction pairs: hand *gesture* and *positive* response, and *passing* and *receiving* of an item. In fact our results show that such micro-actions and reactions cannot be reliably recognized without both sources of information.

Based on these two observations, we address the task using paired egocentric videos recorded by persons A and B to recognize micro-actions or reactions done by person A. Our proposed method works as follows. For each video, we first extract features of first-person POV observations of the self and second-person observations of his/her partner person (each **row** in Figure 2). Features for each person are then collected across videos to provide multiple POV features of the behavior (each **column** in the figure). These features are finally trained individually for A and B and fused to recognize A's micro-actions and reactions.

The main contributions of this work are as follows: (1) we propose the concept of micro-actions and micro-reactions, which are crucial for understanding the dynamics of social interactions; (2) we show that first-person and second-person POV features of two interacting parties are complementary and essential for recognizing micro-actions and reactions; and (3) we construct a new dataset of dyadic interactions comprising more than 1000 pairs of egocentric videos to enable systematic evaluations on micro-action and reaction recognition.

Related Work. First-person vision is one of the emerging topics in computer vision, which greatly affects several applications such as automatic activity summarization [4, 15, 20, 43] and assistive systems [16, 34, 35, 36]. Many studies have used egocentric videos to recognize behaviors of camera wearers, such as action/activity recognition [7, 9, 13, 18, 25, 27, 30], object recognition [10], and gaze estimation [17], where all the visual events in input videos are implicitly assumed to be relevant to the wearer's behaviors. More recently, there has been an interest in understanding group activities recorded in the egocentric videos: *e.g.*, social relationships [2, 3], eye contacts [44, 45], or joint attention [8]. Particularly, Ryoo and Matthies have addressed the problem of recognizing interactions from egocentric videos [29]. They, however, relied on a single video and recognized what a person in the video was doing to a stationary observer. In short, how we can use egocentric videos of two interacting people for recognizing micro-actions and reactions is still unexplored.

Similar to this work, there have been some attempts to use multiple videos but for other purposes. Temporally-aligned egocentric videos can be used for identifying wearers [46] and estimating joint focus of attention [21, 22, 23]. Other work has associated egocentric videos with third-person POV videos (*e.g.*, surveillance videos) for wearer identification [26] and localization [5]. Another relevant task in which multiple videos are used is cross-view action recognition [12, 19, 41], where they focus on the variations in appearances of actions in accordance with the changes in the pose and position of third-person cameras.

To the best of our knowledge, this work is the first to focus on micro-actions and reactions in human-to-human interactions. It is also unlike previous studies that recognize interactions from a third-person POV [6, 11, 28, 37]. We will show that the combination of first-person and second-person POV information enabled by the egocentric videos allows us to recognize various micro-actions and reactions that cannot be well observed in the third-person videos.

2. Our Approach

Suppose that we are given a pair of egocentric videos captured synchronously by person A and his/her partner person B during a dyadic interaction. In each video pair, we assume that person A performs one of several micro-actions and reactions. The goal of this work is to classify these micro-actions/reactions of A from the paired videos.

2.1. Recognition from Paired Egocentric Videos

Our recognition method relies on the multiple POV observations of both persons A and B presented in Figure 2. To this end, we first consider first-person POV features denoted by $\boldsymbol{f}_A, \boldsymbol{f}_B \in \mathbb{R}^{N_{\text{first}}}$ (N_{first} is the number of feature dimensions). These features are extracted from videos recorded by the self to describe a holistic change in the videos such as global motion patterns induced by head motion. We also introduce second-person POV features of each person obtained from the video taken by the other person (*i.e.*, A observed from B's POV and vice versa), $\boldsymbol{f}_{A\leftarrow B}, \boldsymbol{f}_{B\leftarrow A} \in \mathbb{R}^{N_{\text{second}}}$ (N_{second} is the number of feature dimensions). These features should be useful for capturing whole body appearance and motion of the person.

The first-person POV feature $\boldsymbol{f}_A$ and second-person POV feature $\boldsymbol{f}_{A\leftarrow B}$ are then combined to provide multiple POV features of person A. To recognize A's actions and reactions, we define a standard linear decision function to describe the relative importance of the first-person and second-person features:

$$c_A = (\boldsymbol{w}^{(A)}_{\text{first}})^{\mathrm{T}} \boldsymbol{f}_A + (\boldsymbol{w}^{(A)}_{\text{second}})^{\mathrm{T}} \boldsymbol{f}_{A\leftarrow B} + u^{(A)}, \quad (1)$$

where $c_A \in \mathbb{R}$ is a decision score indicating how likely A is to perform a certain action or reaction. $\boldsymbol{w}^{(A)}_{\text{first}} \in \mathbb{R}^{N_{\text{first}}}$, $\boldsymbol{w}^{(A)}_{\text{second}} \in \mathbb{R}^{N_{\text{second}}}$, and $u_A \in \mathbb{R}$ are model parameters describing the importance of each feature; they can be optimized by training any classifiers such as a linear SVM.

Likewise, the multiple POV features for person B are obtained by combining the first-person feature $\boldsymbol{f}_B$ and second-person one $\boldsymbol{f}_{B\leftarrow A}$. We observe that actions or reactions taken by B are often affected by those of A, and thus the features extracted from B's behaviors can be a salient cue to recognize A's actions/reactions. For example, actions of passing an item by A can come with reactions of B receiving the item. Horizontal head rotations of A can stand for a negative response when B is talking to A, while they mean a shift in attention if B is pointing somewhere.

Our proposed method takes advantage of this relationship between A and B by refining A's decision score c_A with B's multiple POV features. Specifically, we introduce another decision function that classifies A's actions and reactions but is learned from $\boldsymbol{f}_B$ and $\boldsymbol{f}_{B\leftarrow A}$:

$$c_B = (\boldsymbol{w}^{(B)}_{\text{first}})^{\mathrm{T}} \boldsymbol{f}_B + (\boldsymbol{w}^{(B)}_{\text{second}})^{\mathrm{T}} \boldsymbol{f}_{B\leftarrow A} + u^{(B)}. \quad (2)$$

Finally, c_A is biased by the score c_B:

$$c'_A = c_A + c_B. \quad (3)$$

This bias can work as follows. To enable classification, we learn functions in Eqs. (1) and (2) for each of several actions and reactions. Even if two micro-reactions (*e.g.*, a negative response and an attention orientation with slight head motion) have similar scores in c_A, the difference in actions by B appears in the score c_B so that we can correct classification results in c'_A.

2.2. First-Person POV Features

In this section, we discuss how various features for first-person action recognition can be used as a first-person POV feature to enable micro-action and reaction recognition.

2.2.1 Egocentric and Object Features

Li *et al.* have focused on actions during hand-manipulation activities (*e.g.*, cooking) [18]. They revealed that effective features included head motion (homography between consecutive frames), hand manipulation points, and object features aligned with dense trajectories [39] around points of gaze and manipulation points, where these features are individually encoded by the Fisher vector (FV) [24].

We expect head motion and object features to work robustly in our dyadic interaction scenarios. Although there may be fewer hand manipulations, the object features could be helpful when large hand motion is apparent from the first-person POV. We therefore adopt the FVs from head motion (**E**) and the combination of the FVs from head motion and object features (**E+O**).

2.2.2 Cumulative Displacement Patterns

Poleg *et al.* have proposed egocentric motion descriptors to enable temporal segmentation of egocentric videos based on activity classes [27]. They rely on cumulative displacement (**CD**) patterns of motion vectors uniformly sampled in video frames, in which we can see long-term trends of egocentric motion in videos. In this work, we aim to use various per-frame features extracted from the CD patterns (such as their slope, motion magnitudes, and *radial projection responses* that Poleg *et al.* presented [27]) to indicate gradual changes of attentional directions.

2.2.3 Pooled Time-Series Encoding

Since the CD features are designed to deal with long-term activities by smoothing motion patterns over time, it may not be optimal to describe short-term cyclic patterns such as head nodding and shaking. We therefore propose to encode the features with the pooled time series (PoT) recently presented by Ryoo *et al.* [30], which we refer to as **PoTCD**. In the encoder, per-frame CD features are first segmented temporally and hierarchically into several shorter patterns. Features in each segment are then encoded by a set of temporal pooling operators such as max/sum pooling and histograms of the positive/negative gradients. This way, our PoTCD features can deal with head motion patterns in detail as well as the gradual changes of attentional directions.

2.3. Second-Person POV Features

We introduce several generic action descriptors for second-person POV features that do not particularly require human detection. These features allow us to capture detailed appearances and motion of people observed from other people and work robustly against significant global motion induced by the head motion of camera wearers.

2.3.1 Improved Dense Trajectory

The improved dense trajectory (**IDT**) [38, 39] is a standard feature descriptor for action recognition used in third-person POV videos [42] as well as egocentric videos [18]. In the IDT, feature points are densely sampled based on the good-feature-to-track [31] and tracked over a short time period (*e.g.*, 15 frames) in accordance with dense optical flow fields. Features such as the histogram of oriented gradients (HOG), the histogram of oriented flows (HOF), and motion boundary histograms (MBH) are then extracted along trajectories and encoded by the FV.

We believe that the IDT is well suited to describe people from a second-person POV since it can extract relevant motion of the people without tracking them explicitly.

2.3.2 Two-Stream Convolutional Networks

Instead of hand-crafted features like the dense trajectory, Simonyan and Zisserman [32] learned feature representations and action classifiers in a convolutional neural network (CNN). Particularly, they introduced two-stream CNN (**TCNN**) where two CNNs individually learned the appearances and motions over a short period (*e.g.*, 20 frames).

A CNN trained on a relevant dataset (*e.g.*, action recognition datasets such as UCF101 [33] and HMDB51 [14]) can also be used as a feature descriptor. In this study, we use some mid-level convolution outputs drawn from the two CNNs and encoded them by the FV to serve as second-person POV features. For input motion sequences, we compute local motion vectors by subtracting global motion displacements from original optical flow fields.

2.3.3 Trajectory-Pooled Convolutional Descriptors

While the **TCNN** can provide rich information on both of the appearances and motion of people in videos, it encodes all the events occurring in the videos regardless of whether they belong to foregrounds (people) or backgrounds. To resolve this problem, we further pool TCNN features along dense trajectories as proposed by Wang *et al.* [40] (which they refer to as **TDD**). Features extracted in this way can be limited to relevant events where trajectories appear.

3. Experiments

We first systematically evaluate how the features introduced in the previous section can work on the task of detecting specific micro-actions and reactions observed during dyadic interactions in Sections 3.3 and 3.4. We also investigate how our method can classify micro-actions and reactions in Sections 3.5 and 3.6. Implementation details are described in the appendix.

3.1. Paired Egocentric Video Dataset

Among the datasets of egocentric videos released to date, only a few include interaction scenes. JPL interaction dataset [29] and EGO-GROUP dataset [2, 3] comprise only one POV video for each interaction scene. While the first-person social interactions dataset [8], CMU first-person video dataset [21, 22], and ego-surfing dataset [46] provide egocentric videos of multiple people, none has enough interaction sequences to enable supervised learning of micro-actions and reactions from multiple POV observations.

In this work, we present a new video dataset named *Paired Egocentric Video (PEV)* dataset, a large collection of paired egocentric videos recorded during dyadic human-to-human interactions. The dataset contains 1226 pairs of videos in total, each of which includes a single micro-action or reaction pattern of a person regarded as target person A (see Figure 4 for examples). All video pairs were selected from several continuous recordings of face-to-face conversations. There were six subjects wearing different clothes in eight different everyday environments such as a cafeteria and an office. Actions and reactions in the data have variability in motion and appearance since we did not particularly instruct subjects on how and when to perform actions or reactions during the recordings. We did however inform each subject of the following seven action/reaction types that we aimed to collect.

1. **Pointing (154 samples)**: Pointing to a certain location, an item, or person B to initiate interaction, which is followed by B's reactions such as orienting of attention and positive or negative responses.

2. **Attention (97 samples)**: Orienting attention with slight head motion to what is pointed to by B.

3. **Positive (159 samples)**: Responding positively by widely or subtly nodding and/or by laughing with body motion to B's pointing or gesture.

4. **Negative (40 samples)**: Responding negatively by shaking or slightly cocking one's head and/or crossing arms to B's pointing or gesture.

5. **Passing (150 samples)**: Initiating or finishing passing an item to B in order to exchange it.

6. **Receiving (143 samples)**: Initiating or finishing receiving what B is trying to pass.

7. **Gesture (168 samples)**: Doing head and/or hand gestures to converse with B, which can be followed by B's gesture and positive or negative responses.

Note that the remaining 315 pairs in the dataset contain non-interaction patterns where person A is just moving that are irrelevant to the current context of interactions: *e.g.*, placing an item on a table or looking at a certain location to which person B did not particularly point. Each video has 90 frames (1.5 seconds at 60 fps) and the spatial resolution of 320x180, where the 30th frame of each video was adjusted to the onset of actions and reactions of A.

3.2. Evaluation Scheme

Since the six subjects formed three pairs in the dataset, we conducted a three-fold cross validation by splitting the data into subsets on based on the pairs. We trained the decision functions in Eq. (1) and Eq. (2) by using two training subsets and evaluated performance with one testing subset.

In Sections 3.3 and 3.4, we evaluate detection performance by the area under the receiver-operating characteristic curve (AUC score) computed from decision scores (*e.g.*, c_A, c'_A) and binary ground-truth labels (1 for the correct actions/reactions and 0 otherwise) collected from all three tests. In the classification task in Sections 3.5 and 3.6, we further normalize the decision scores to have zero-mean and unit-variance for each action/reaction and compare them for each sample to find the most probable one. Average accuracies over all the actions and reactions are calculated for the classification performance.

3.3. Comparison among First-Person and Second-Person POV Features

We first focused on the use of single egocentric videos and compared detection performance for first-person POV features (**E** [18], **E+O** [18], **CD** [27], **PoTCD** [27, 30]) and second-person ones (**IDT** [39], **TCNN** [32] and **TDD** [40]) of target person A. To this end, we performed detection based on A's decision score c_A where the function in Eq. (1) was learned from either $\boldsymbol{f}_A$ or $\boldsymbol{f}_{A\leftarrow B}$.

Table 1(1) shows AUC scores using first-person features. Overall, these features worked well for detecting reactions with head motion such as *attention*, *positive*, and *negative*. **CD** had a limited performance as it was not well suited to cyclic motion such as nodding. **E+O** performed better on *receiving* when large hand motion was made at the center of first-person POV clearly and captured by object features.

Among the second-person POV features described in Table 1(2), **IDT** performed better when the actions and reactions involved hand motion such as *pointing*, *passing*, *receiving*, and *gesture*. It worked particularly well on *receiving* since people often received items in front of their body that were clearly visible from the second-person POV. On the other hand, **TCNN** and **TDD** provided inferior scores. We found that the location where people appeared in egocentric videos often changed drastically over time due to significant head motion of camera wearers. As CNNs used in these two methods encoded appearances and motion at every fixed location at predefined intervals (20 frames), resultant features often became irrelevant when the location of people changed in a time shorter than the interval.

3.4. Combining Multiple POV Features

We then investigated how the performance was improved by combining multiple POV features. In what follows, we pick out the four features producing AUC scores over 0.7 in the previous section: **E**, **E+O**, **PoTCD**, and **IDT**.

Table 1(3) shows results from the combination of first-person POV and second-person POV features of target person A. Specifically, we evaluated A's decision score c_A learned from both $\boldsymbol{f}_A$ and $\boldsymbol{f}_{A\leftarrow B}$. All the combined methods performed well regardless of whether actions and reactions came with head and/or hand motion, meaning that first-person and second-person features worked complementarily in the methods. Furthermore, Table 1(4) confirms that the combination of multiple POV features of person A and those of B performed the best. In the **Proposed** method, we evaluated decision score c'_A in Eq. (2) where the feature **PoTCD** was used for $\boldsymbol{f}_A, \boldsymbol{f}_B$ and **IDT** for $\boldsymbol{f}_{A\leftarrow B}, \boldsymbol{f}_{B\leftarrow A}$. These results indicate that first-person and second-person POV observations of two people are essential for recognizing micro-actions and reactions.

To analyze the effect of using paired egocentric videos in more detail, we implemented some degraded versions of

Table 1. AUC scores on the detection task. (1) First-person POV features of target person A. (2) Second-person POV features of A. (3) Combinations of first-person and second-person POV features of A. (4) Combinations of multiple POV features of persons A and B.

		Pointing	Attention	Positive	Negative	Passing	Receiving	Gesture	Average
(1) First-person POV features of A	**E** [18]	0.65	0.77	0.91	0.88	0.64	0.78	0.73	0.76
	E+O [18]	0.74	0.77	0.94	0.73	0.71	0.85	0.69	0.77
	CD [27]	0.64	0.62	0.58	0.56	0.71	0.71	0.56	0.63
	PoTCD [27, 30]	0.70	0.66	0.94	0.84	0.69	0.74	0.63	0.74
(2) Second-person POV features of A	**IDT** [39]	0.74	0.71	0.67	0.59	0.81	0.93	0.78	0.75
	TCNN [32]	0.59	0.58	0.55	0.58	0.54	0.67	0.60	0.59
	TDD [40]	0.63	0.70	0.61	0.51	0.68	0.79	0.63	0.65
(3) Multiple POV features of A	**E+IDT**	0.77	0.73	0.86	0.81	0.82	0.92	0.79	0.81
	E+O+IDT	0.80	0.78	0.95	0.77	0.83	0.95	0.78	0.84
	PoTCD+IDT	0.79	0.78	**0.96**	0.89	0.84	0.93	0.80	0.86
(4) Multiple POV features of A and B	**Degraded-A**	0.82	0.76	**0.96**	0.86	0.56	0.95	0.69	0.84
	Degraded-B	0.73	0.72	0.67	0.61	0.82	0.94	0.78	0.75
	Proposed	**0.85**	**0.83**	**0.96**	**0.91**	**0.89**	**0.97**	**0.82**	**0.89**

Table 2. Classification accuracies on the JPL dataset [29].

$\boldsymbol{f}_B$	$\boldsymbol{f}_{A\leftarrow B}$	**Degraded-B**
0.61	0.70	**0.75**

Proposed given only one of the two videos. In **Degraded-A**, we used the video recorded by person A to learn c_A from only $\boldsymbol{f}_A$ and c_B from $\boldsymbol{f}_{B\leftarrow A}$. On the other hand, **Degraded-B** accepted the video recorded by B and adopted $\boldsymbol{f}_{A\leftarrow B}$ in c_A and $\boldsymbol{f}_B$ in c_B. Note that **Degraded-B** has the same conditions as the work of Ryoo and Matthies [29]: only videos including a target person from the second-person POV were available. The decreased performance of these methods in Table 1(4) indicates the necessity of observing both videos.

3.5. Classifying Micro-Actions and Reactions

We finally investigated how our method could classify different micro-actions and reactions. We picked out **PoTCD** in Table 1(1), **IDT** in (2), **PoTCD** in (3), and **Proposed** in (4) as they provided good detection performance.

Figure 3 describes confusion matrices. As *passing*, *receiving*, and *gesture* often appeared subtly in front of one's body, they were difficult to classify where only the first-person feature **PoTCD** was given. On the other hand, **IDT** could classify them while it was less discriminative on micro-reactions with subtle head motion such as *attention*, *positive*, and *negative*. We confirmed that **PoTCD+IDT** inherited the advantages of first-person and second-person features. **Proposed** further improved the performance on *attention*, *positive*, and *receiving* even when they came with small motions because these actions/reactions of A often induced different behavior of B.

Figure 4 presents some visual examples of classification results together with dense trajectories [39] and cumulative displacement patterns [27]. When hand motion was distinct in both of the first-person and second-person POVs (*e.g.*, the *pointing* action annotated by the arrows in example (1)), all the methods were able to predict a correct action. Some micro-actions and reactions were observed with the combination of head and hand motions. These motions were not always large enough, as the *pointing* by person A could not be seen well in his/her first-person POV in example (2), or the *nodding* was very slight in the second-person POV in (3). Even for such cases, these two POV sources complementarily worked in **PoTCD+IDT** and **Proposed**. People seen in the second-person POV were often partially occluded especially when they were focusing on objects of interest as annotated in example (4). As the second-person features in Section 2.3 did not rely on human detection, our approach was robust against such cases.

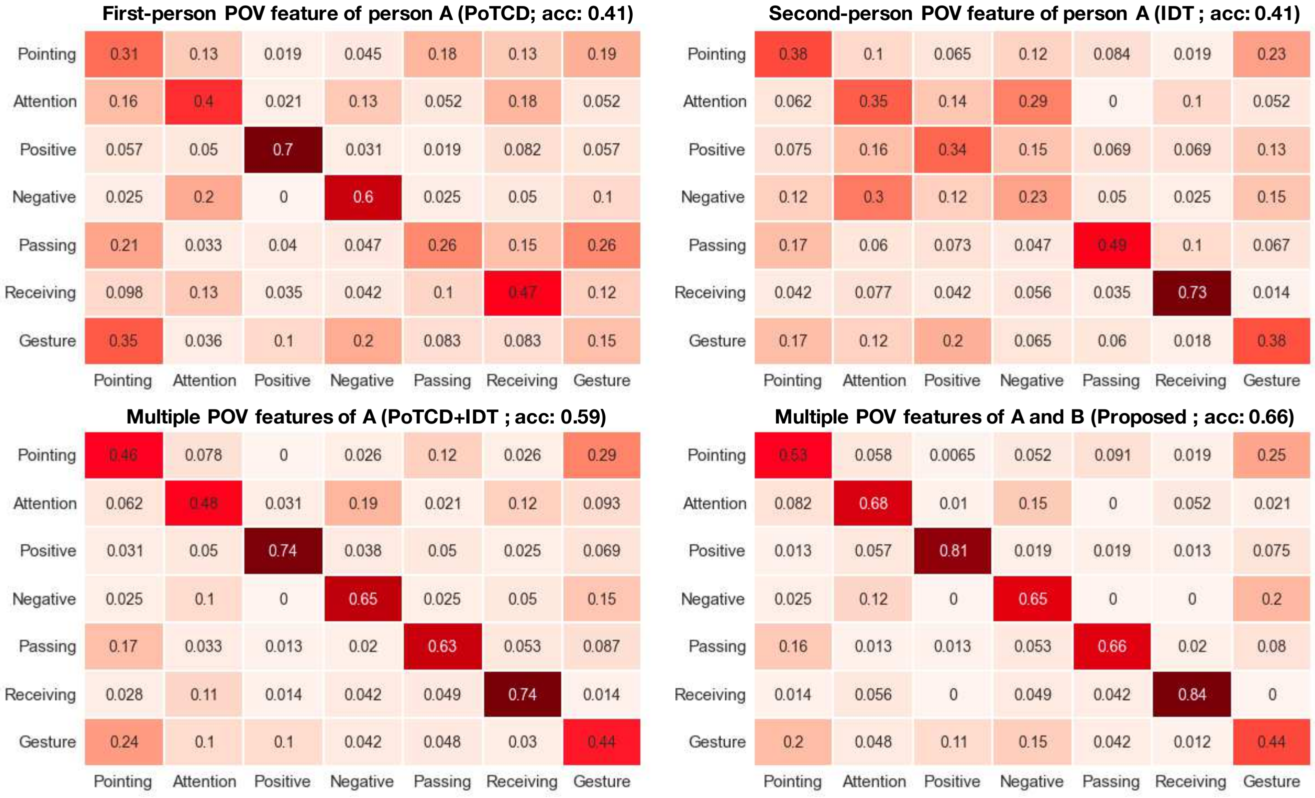
We also found some temporal structures between the head motion of two people. For instance, the cumulative displacement patterns in example (5) illustrate that the head motion of A induced by the shift in *attention* was followed by the head motion of B to share attention. Similarly, mutual head motion was found when responding *negative*ly as annotated in example (6). **Proposed** could classify such micro-reactions successfully by exploiting both actions and reactions of the two persons.

Gesture was the most difficult class to recognize even for our method (examples (7) and (8)). As annotated in the examples, *gesture* required both first-person and second-person features since it often came with head and hand motion. This motion was however sometimes similar to other actions or reactions such as *pointing*, *positive*, and *negative*.

3.6. Evaluation on the JPL Dataset

We also evaluated the classification performance of our method on the JPL dataset [29][1]. It includes seven activities of a target person such as handshakes, hugs, and punches, some of which lasted longer (several seconds) than ours. As this dataset is composed of only the egocentric videos of a stationary observer (standing for person B), we compared **Degraded-B** against its degraded versions using either first-person feature $\boldsymbol{f}_B$ or second-person one $\boldsymbol{f}_{A\leftarrow B}$ to see the

[1] http://michaelryoo.com/jpl-interaction.html



First-person POV feature of person A (PoTCD; acc: 0.41)

	Pointing	Attention	Positive	Negative	Passing	Receiving	Gesture
Pointing	0.31	0.13	0.019	0.045	0.18	0.13	0.19
Attention	0.16	0.4	0.021	0.13	0.052	0.18	0.052
Positive	0.057	0.05	0.7	0.031	0.019	0.082	0.057
Negative	0.025	0.2	0	0.6	0.025	0.05	0.1
Passing	0.21	0.033	0.04	0.047	0.26	0.15	0.26
Receiving	0.098	0.13	0.035	0.042	0.1	0.47	0.12
Gesture	0.35	0.036	0.1	0.2	0.083	0.083	0.15

Second-person POV feature of person A (IDT ; acc: 0.41)

	Pointing	Attention	Positive	Negative	Passing	Receiving	Gesture
Pointing	0.38	0.1	0.065	0.12	0.084	0.019	0.23
Attention	0.062	0.35	0.14	0.29	0	0.1	0.052
Positive	0.075	0.16	0.34	0.15	0.069	0.069	0.13
Negative	0.12	0.3	0.12	0.23	0.05	0.025	0.15
Passing	0.17	0.06	0.073	0.047	0.49	0.1	0.067
Receiving	0.042	0.077	0.042	0.056	0.035	0.73	0.014
Gesture	0.17	0.12	0.2	0.065	0.06	0.018	0.38

Multiple POV features of A (PoTCD+IDT ; acc: 0.59)

	Pointing	Attention	Positive	Negative	Passing	Receiving	Gesture
Pointing	0.46	0.078	0	0.026	0.12	0.026	0.29
Attention	0.062	0.48	0.031	0.19	0.021	0.12	0.093
Positive	0.031	0.05	0.74	0.038	0.05	0.025	0.069
Negative	0.025	0.1	0	0.65	0.025	0.05	0.15
Passing	0.17	0.033	0.013	0.02	0.63	0.053	0.087
Receiving	0.028	0.11	0.014	0.042	0.049	0.74	0.014
Gesture	0.24	0.1	0.1	0.042	0.048	0.03	0.44

Multiple POV features of A and B (Proposed ; acc: 0.66)

	Pointing	Attention	Positive	Negative	Passing	Receiving	Gesture
Pointing	0.53	0.058	0.0065	0.052	0.091	0.019	0.25
Attention	0.082	0.68	0.01	0.15	0	0.052	0.021
Positive	0.013	0.057	0.81	0.019	0.019	0.013	0.075
Negative	0.025	0.12	0	0.65	0	0	0.2
Passing	0.16	0.013	0.013	0.053	0.66	0.02	0.08
Receiving	0.014	0.056	0	0.049	0.042	0.84	0
Gesture	0.2	0.048	0.11	0.15	0.042	0.012	0.44

Figure 3. Confusion matrices and average accuracies of the classification task on the PEV dataset.

effectiveness of combining multiple POV features. We followed the same protocol as Ryoo and Matthies [29] and repeated two-fold cross validations 100 times.

As shown in Table 2, we found that the combination of first-person and second-person features in **Degraded-B** performed the best. Note that the method of Ryoo and Matthies [29] performs better (0.896 as a classification accuracy) by incorporating structured prediction tailored to long-term activities with multiple sub-events. Future work will be to extend our method to cope with multiple action and reaction sequences.

3.7. Limitations

One current limitation of our method is that it only considers behaviors of two people taking place in the same time period. Recognizing actions and reactions with a large amount of delay will require a structured prediction [28, 30, 37]. In addition, we currently focus on only two-person scenarios. To generalize our work to deal with group interactions where more than two people are present, wearer identification [26, 46] will be necessary to obtain a second-person POV observation of specific persons.

4. Conclusions

We have introduced the task of recognizing micro-actions and reactions in dyadic human-to-human interactions. The key finding of our work is that the micro-actions and reactions can be best recognized by utilizing first-person and second-person POV features of two interacting people. Understanding social interaction dynamics by recognizing micro-actions and reactions will impact several first-person vision tasks such as video summarization of social events and assistive systems, and also raise new problems such as wearer identification in crowded scenes and modeling of group interaction dynamics.

A. Implementation Details

We adopted a linear SVM for the decision functions in Eq. (1) and Eq. (2) and trained them via stochastic gradient descent as it performed the best. As **CD** [27] features were obtained per frame, we computed decision scores for each frame and averaged them over time. In **PoTCD** [27, 30], we used only three pyramids (split into one, two, and four segments) as we worked on short video clips.

On **E**, **E+O** [18], **IDT** [39], **TCNN** [32] and **TDD** [40], we learned some additional models for the FV in the training datasets: the principal component analysis (PCA) to

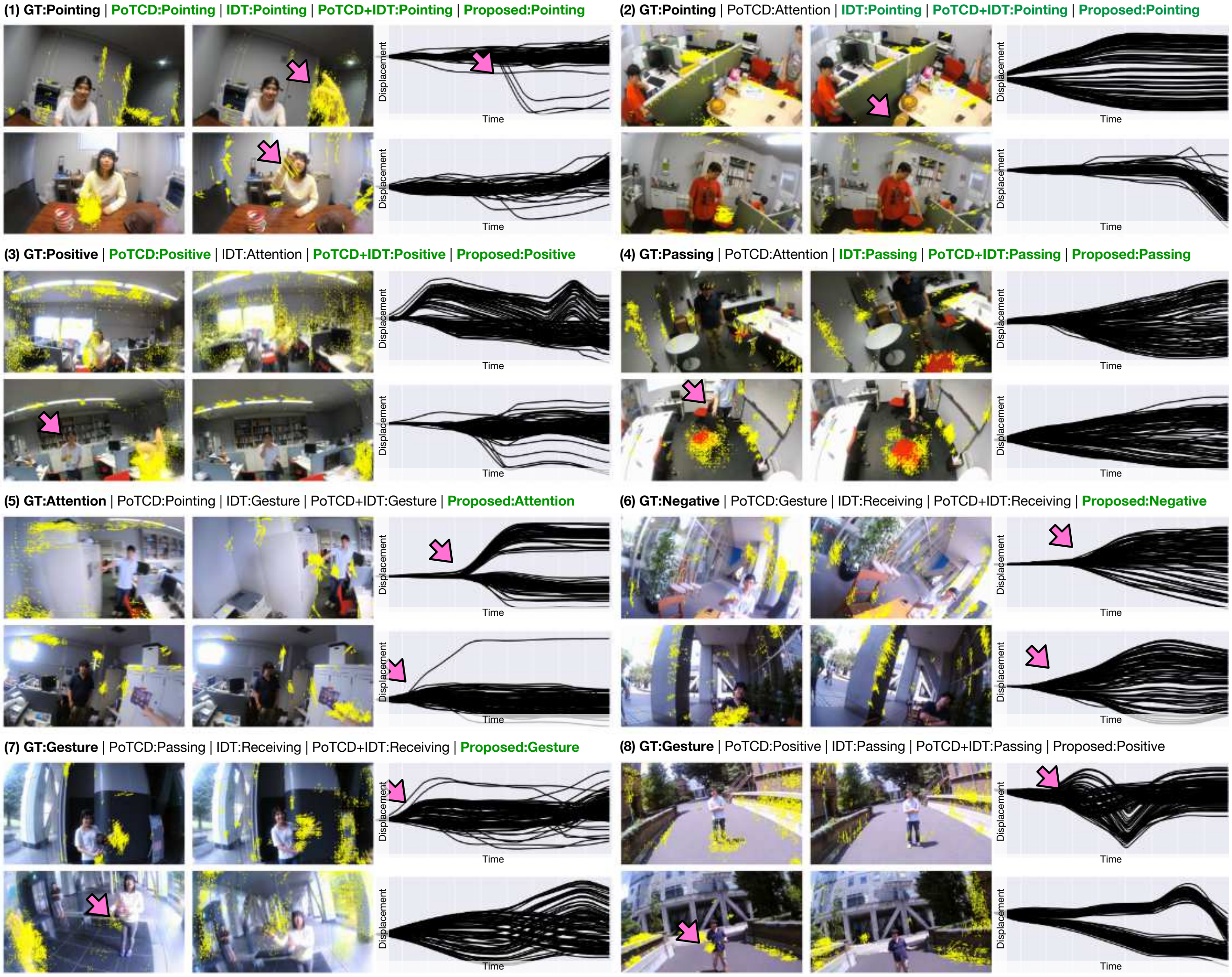


Figure 4. Our proposed method working on the PEV dataset. The first row of each example shows 40th and 70th frames of the video recorded by target person A as well as its cumulative displacement patterns [27] (motion vectors uniformly sampled in video frames and accumulated over time) that are encoded by pooled time-series [30] in our proposed method. Dense trajectories [39] are visualized by the yellow arrows in each video frame. The second row of each example provides the same visualization but for the video recorded by person B. Titles describe classification results (correct classifications are highlighted in green) as well as the ground-truth label. Micro-actions and reactions annotated by the pink arrows are discussed in Section 3.5.

perform a dimensionality reduction on features and the Gaussian mixture model (GMM) to generate the FV. We followed the original papers [18, 39, 40] to determine the number of components for the PCA and GMM. The PCA with half the number of original feature dimensions and the GMM with 64 components were used for **E** and **E+O**, and the PCA with 64 components and the GMM with 256 components were used for **IDT**, **TCNN**, and **TDD**. All the FVs were further applied the power and L2 normalizations [24].

We used the code available on the web[2] for dense trajectories in **E+O**, **IDT** and **TDD**. As hand manipulations were barely found in the PEV dataset, in **E+O** we extracted the object features along all the trajectories. We adopted the CNNs trained by Wang *et al.* [40] for **TCNN** and **TDD**. Based on the results from each convolution layer in the work of Wang *et al.* [40], we concatenated the outputs of the conv4 layer of spatial CNN and the conv3 layer of temporal CNN as second-person features.

Acknowledgments

This research was supported by CREST, JST and Kayamori Foundation of Informational Science Advancement.

[2] https://lear.inrialpes.fr/people/wang/improved_trajectories

References

[1] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara. Head pose estimation in first-person camera views. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 1–6, 2014. 1

[2] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara. Understanding social relationships in egocentric vision. *Pattern Recognition*, 48(12):4082 – 4096, 2015. 1, 2, 4

[3] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 594–599, 2014. 1, 2, 4

[4] I. Arev, H. S. Park, Y. Sheikh, J. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics*, 33(4):81:1–81:11, 2014. 2

[5] V. Bettadapura, I. Essa, and C. Pantofaru. Egocentric field-of-view localization using first-person point-of-view devices. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015. 2

[6] C.-W. Chen, R. Ugarte, C. Wu, and H. Aghajan. Discovering social interactions in real work environments. In *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG)*, pages 933–938, 2011. 3

[7] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 407–414, 2011. 1, 2

[8] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1226–1233, 2012. 1, 2, 4

[9] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 314–327. Springer-Verlag, 2012. 1, 2

[10] A. Fathi, X. Ren, and J. Rehg. Learning to recognize objects in egocentric activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3281–3288, June 2011. 2

[11] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 489–504, 2014. 3

[12] I. Junejo, E. Dexter, I. Laptev, and P. Perez. View-independent action recognition from temporal self-similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(1):172–185, 2011. 2

[13] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3241–3248, 2011. 1, 2

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2556 – 2563. 4

[15] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346 – 1353, 2012. 2

[16] T.-S. Leung and G. Medioni. Visual navigation aid for the blind in dynamic environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 153 – 158, 2014. 2

[17] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3216–3223, 2013. 2

[18] Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 287 – 295, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[19] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3209–3216, 2011. 2

[20] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721, 2013. 2

[21] H. S. Park, E. Jain, and Y. Sheikh. 3d social saliency from head-mounted cameras. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 1–9, 2012. 2, 4

[22] H. S. Park, E. Jain, and Y. Sheikh. Predicting primary gaze behavior using social saliency fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3503 – 3510, 2013. 2, 4

[23] H. S. Park and J. Shi. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4777–4785, 2015. 2

[24] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 143–156, 2010. 3, 8

[25] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2847–2854, 2012. 1, 2

[26] Y. Poleg, C. Arora, and S. Peleg. Head motion signatures from egocentric videos. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 1–15, 2014. 2, 7

[27] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2537–2544, 2014. 1, 2, 3, 4, 5, 6, 7, 8

[28] M. Ryoo and J. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1593–1600, 2009. 3, 7

[29] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2737, 2013. 1, 2, 4, 6, 7

[30] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 896 – 904, 2015. 1, 2, 4, 5, 6, 7, 8

[31] J. Shi and C. Tomasi. Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593 – 600, 1994. 4

[32] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 568–576, 2014. 4, 5, 6, 7

[33] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. Technical report, 2012. 4

[34] T. J. J. Tang and W. H. Li. An assistive eyewear prototype that interactively converts 3d object locations into spatial audio. In *Proceedings of the ACM International Symposium on Wearable Computers (ISWC)*, pages 119–126, 2014. 2

[35] R. Templeman, M. Korayem, D. Crandall, and A. Kapadia. Placeavoider: Steering first-person cameras away from sensitive spaces. In *Proceedings of the Annual Network and Distributed System Security Symposium (NDSS)*, 2014. 2

[36] Y. Tian, Y. Liu, and J. Tan. Wearable navigation system for the blind people in dynamic environments. In *Proceedings of the Cyber Technology in Automation, Control and Intelligent Systems (CYBER)*, pages 153 – 158, 2013. 2

[37] A. Vahdat, B. Gao, M. Ranjbar, and G. Mori. A discriminative key pose sequence model for recognizing human interactions. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1729–1736, 2011. 3, 7

[38] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176, 2011. 4

[39] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3551 – 3558, 2013. 2, 3, 4, 5, 6, 7, 8

[40] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4305 – 4314, 2015. 4, 5, 6, 7, 8

[41] L. Xia, C.-C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 20–27, 2012. 2

[42] H. Xu, Q. Tian, Z. Wang, and J. Wu. A survey on aggregating methods for action recognition with dense trajectories. *Multimedia Tools and Applications*, pages 1–17, 2015. 4

[43] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2235–2244, 2015. 2

[44] 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 ACM Conference on Ubiquitous Computing (UbiComp)*, pages 699–704, 2012. 1, 2

[45] Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. M. Rehg. Detecting bids for eye contact using a wearable camera. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1 – 8, 2015. 1, 2

[46] R. Yonetani, K. M. Kitani, and Y. Sato. Ego-surfing first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5445–5454, 2015. 2, 4, 7