Conversation activity recognition using interaction video sequences in pedestrian groups

Wataru Ganaha, Takumi Ozaki, Michiko Inoue, and Masashi Nishiyama[0000−0002−5964−3209]

Graduate School of Engineering, Tottori University, 101 Minami 4-chome, Koyama-cho, Tottori, 680-8550, Japan nishiyama@tottori-u.ac.jp

Abstract. We introduce a method for recognizing conversation activity in a group of people walking outdoors using a color video sequence acquired from a camera. Many methods have been developed to recognize whether people are walking together or talking together in a color video sequence. However, a method has yet to be proposed to recognize conversation activity in a pedestrian group walking outdoors. In this paper, we design a feature extraction approach for conversation activity recognition using physical body interactions caused by pedestrians' conversations. Our method generates an interaction video sequence in a virtual space using a temporal posture signal and a temporal walking position signal that represent pedestrians' body interactions. Our method uses the interaction video sequence as an informative and visible feature to determine a conversation activity label. The experimental results showed that our interaction video sequence recognized conversation activity more accurately than alternative techniques that use the appearance of the body regions of a pedestrian group or time-series changes of the posture and walking position among pedestrians.

Keywords: Conversation activity recognition · Pedestrian groups · Human body interaction.

1 Introduction

A demand exists for technology that can automatically recognize human interactions within a group of people walking outdoors. In this study, we focus on conversation activity as one form of human interaction in a pedestrian group. We define conversation activity as whether a conversation is occurring within a pedestrian group and whether the conversation is active or inactive. One possible application of conversation activity recognition is marketing in a scenario in which many pedestrian groups are walking in the aisles of a shopping mall. Figure 1 shows an example of the application. By comparing the number of pedestrian groups engaged in active conversation between visitors that are arriving and leaving, it may be possible to determine whether visitors are satisfied with their visit.

We consider what feature can be used to recognize conversation activity in a pedestrian group. A possible feature is the chronological change of speech sounds,

Fig. 1: We assume that an application for conversation activity recognition exists. This application can determine whether pedestrian groups are satisfied with their visit by comparing the number of pedestrian groups engaged in active conversation between visitors that are arriving and leaving.

such as the timing of pedestrians' utterances and the inflection of pedestrians' voices. However, because we target a group of pedestrians walking outdoors, it is difficult to use a microphone for voice sensing for each pedestrian. Instead, we consider using a color video sequence acquired from a surveillance camera as a feature that represents a human body interaction performed in a pedestrian group. We assume that the time-series changes in gestures performed by each pedestrian, and the time-series changes in pedestrians' body orientation and walking position, provide a visible and informative feature for conversation activity recognition. When analyzing speech among people [12], it is well known that gestures, that is, movements produced by the body in response to speech, are helpful. Regarding the analysis of pedestrian group behavior [23] and the development of the group detection method [2], it is well known that body orientation and the walking position, which are interrelated among pedestrians that belong to one group, are helpful. In this study, our definition of physical body interaction consists of gestures, pedestrians' body orientation, and pedestrians' walking position.

We consider how to design a method to recognize conversation activity using the body interaction feature in a video sequence. To the best of our knowledge, a method has yet to be proposed to recognize conversation activity in a pedestrian group. Instead, we survey existing methods for recognizing the presence or absence of body interaction in a pedestrian group, such as whether the pedestrians are walking together or talking together in a video sequence. These existing methods can be divided into two main categories. The first category contains methods [5, 24, 2, 18, 19] that detect the presence or absence of a pedestrian group. The second category contains methods [9, 17, 13] that recognize whether people in a group are talking together, given that a pedestrian group has been detected. More recently, methods [3, 6, 15] have emerged that detect the presence or absence of a pedestrian group and simultaneously recognize the presence or absence of conversations within that group. However, even when these existing

methods are applied, it is impossible to recognize whether the conversation is active or inactive in a pedestrian group.

In this paper, we propose a novel method for recognizing conversation activity in a pedestrian group by extracting an interaction video sequence as a feature, which has high recognition accuracy and can be visually confirmed by human observers. Our method generates an interaction video sequence in a pedestrian group using a temporal posture signal and temporal walking position signal estimated from a color video sequence. By applying this interaction video sequence to the class classification network, our method determines a conversation activity label: active conversation, inactive conversation, or no conversation. The active conversation label indicates the state in which the pedestrian group is having a lively conversation on topics of mutual interest. The inactive conversation label indicates the state in which the group is not having a lively conversation on topics of no interest. The no conversation label indicates the state in which no conversation is occurring.

The salient contributions of this paper are as follows:

- We extract an informative feature using an interaction video sequence rendered in a virtual space by fixing the viewpoint position of the virtual camera in front of a pedestrian group.
- We design a visible feature that allows human observers to directly see physical body interaction performed in a pedestrian group.
- On an originally collected outdoor pedestrian dataset of 624 video sequences in 52 groups, we demonstrated that our interaction video sequences achieved high accuracy in conversation activity recognition.

From the experimental results, we confirmed that our method using an interaction video sequence recognized conversation activity more accurately than using color video sequences of pedestrian body regions or using a temporal posture and walking position signal.

2 Method for recognizing conversation activity

2.1 Overview

In this paper, we assume that body interaction arising from conversation activity among pedestrians is represented explicitly by time-series signals of the posture and walking position. Figure 2 shows an overview of our method. In the following, we describe the procedure in our method.

P1. Body region estimation:
We estimate the region that represents the body of a pedestrian at each time point in a color video sequence acquired from a surveillance camera.

P2. Temporal posture signal estimation:
We estimate a temporal posture signal from the appearance of a pedestrian body region at each time point. Specifically, we use the three-dimensional (3D) human body model to extract a time-series signal that represents only

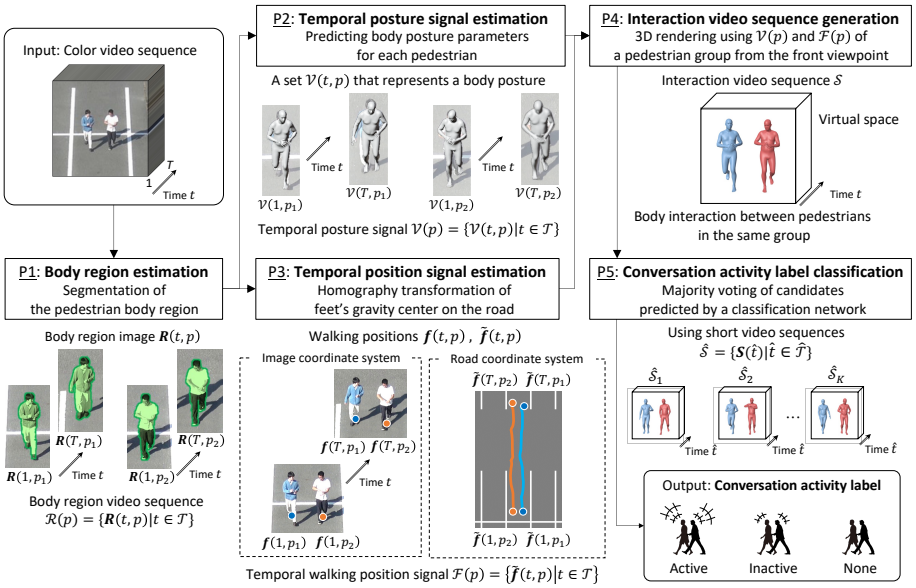

Fig. 2: Overview of our method for conversation activity recognition. We estimate body region images from a color video sequence acquired from a camera in P1. Our method obtains a temporal posture signal and temporal walking position signal that represents the body interaction caused by the conversation in P2 and P3. We generate an interaction video sequence in a virtual space for extracting an informative and visible feature in P4 and determine a conversation activity label using the interaction video sequence in P5.

the posture change of each pedestrian. Using this signal, we represent posture changes in gestures and body orientation while pedestrians engage in conversation in a group.

P3. Temporal walking position signal estimation:
We estimate a temporal walking position signal by calculating the feet's center of gravity from the pedestrian's body region at each time point. Specifically, we estimate the center of gravity of the feet's contour for each pedestrian and determine the walking position on the road surface by applying a homography transformation. Using this signal, we extract temporal changes that represent positional relationships in a conversation in a group.

P4. Interaction video sequence generation:
Our method generates an interaction video sequence of a pedestrian group using 3D rendering with a temporal posture signal of P2 and temporal walking position signal of P3. By always fixing the virtual camera viewpoint in front of the pedestrian group, we extract a feature that can capture the body interaction that effectively recognizes the conversation activity label.

We also design a feature that allows human observers to visually and temporally confirm physical body interactions in a group.

P5. Conversation activity label classification:
We determine the conversation activity label using a classification network for an interaction video sequence of P4. We use three conversation activity labels: active conversation, inactive conversation, and no conversation. We explain the details of these labels in Section 3.2. We generate multiple short video sequences from a single interaction video sequence and output multiple candidate labels from the classification network using these short video sequences. A majority vote among these candidate labels determines the final conversation activity label.

In the following sections, we describe each procedure in detail.

2.2 Body region estimation

In procedure P1, we estimate the pedestrian body region from a video sequence acquired from a camera. The body region video sequence $\mathcal{R}(p)$ that consists of pedestrian body pixels and surrounding background pixels is expressed as

$$\mathcal{R}(p) = \{\boldsymbol{R}(t,p) \mid t \in \mathcal{T}\}, \tag{1}$$

where $\boldsymbol{R}(t,p)$ is the body region image of each pedestrian p at time point t, $\mathcal{T}$ is a set that consists of the times when the images were acquired, and T is the total number of times that belong to the set $\mathcal{T}$. T also represents the length of time from when a pedestrian enters the camera's field of view until the pedestrian leaves. Note that $\boldsymbol{R}(t,p)$ consists of a pedestrian body region and the background region surrounding it. $\boldsymbol{R}(t,p)$ stores a mask, whether each pixel belongs to the body or background region, and the RGB value of each pixel. We use Mask R-CNN [7], which is internally called from within PHALP [16], at each time point to estimate the pedestrian body region. PHALP is a body posture and shape estimation method, as described in the next section. This method also performs pedestrian tracking and determines each pedestrian p of $\boldsymbol{R}(t,p)$.

2.3 Temporal posture signal estimation

In procedure P2, we estimate a temporal posture signal from the body region video sequence $\mathcal{R}(p)$ to represent the changes of gestures and body orientation in a conversation among pedestrians. First, we estimate the pedestrian's posture from the body region image $\boldsymbol{R}(t,p) \in \mathcal{R}(p)$. The posture is denoted by $\mathcal{V}(t,p)$, a set of 3D vertices $\boldsymbol{v}(t,p)$ on the pedestrian's body surface, and their adjacent vertices. A temporal posture signal $\mathcal{V}(p)$ is expressed as

$$\mathcal{V}(p) = \{\mathcal{V}(t,p) \mid t \in \mathcal{T}\}. \tag{2}$$

In this study, to estimate $\mathcal{V}(t,p)$, which represents the posture changes, we apply PHALP [16] described in the previous section. PHALP is a method for tracking people in monocular movies by predicting their future 3D representations.

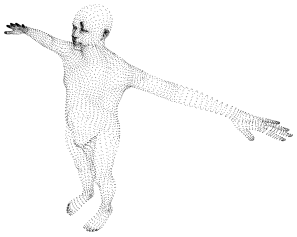

Fig. 3: Examples of vertices on the body surface.

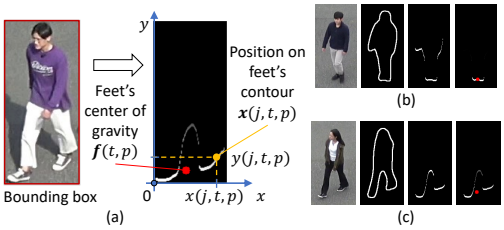

Fig. 4: Parameters used to calculate position $\boldsymbol{f}(t,p)$ of the feet's center of gravity.

This method involves estimating temporal models for the 3D pose, position, and appearance and using these models for probabilistic matching and updating tracklets. PHALP uses SMPL [11], which is a 3D human body model, to represent the posture and body shape parameters. The posture parameters are specifically expressed as a rotation matrix at the 23 joint points of the human body and a rotation matrix over the whole body. Using the estimated posture parameters and the standard body shape parameters, we generate a set $\mathcal{V}(t,p)$ that consists of 6,890 vertices $\boldsymbol{v}(t,p)$ on the body surface and their adjacent vertices. Figure 3 shows examples of vertices on the body surface.

When estimating a temporal posture signal, outliers in the time direction often occur suddenly. We detect outliers by applying a Hampel filter to the time-series signal of the 3D vertex $\boldsymbol{v}(t,p) \in \mathcal{V}(t,p)$. Then we interpolate the posture parameters at the time of the outlier using the nearest neighbor technique from the values at the surrounding time.

2.4 Temporal walking position signal estimation

In P3, we estimate a temporal walking position signal on a road surface to represent the pedestrian's positional relationship caused by the conversation. First, in the body region image $\boldsymbol{R}(t,p) \in \mathcal{R}(p)$, our method estimates the position of the feet's center of gravity $\boldsymbol{f}(t,p)$ in the image coordinate system. Next, by applying a homography transformation to convert the image coordinate system to the road surface coordinate system, our method obtains the walking position $\tilde{\boldsymbol{f}}(t,p)$. The temporal walking position signal $\mathcal{F}(p)$ is expressed as

$$\mathcal{F}(p) = \{\tilde{\boldsymbol{f}}(t,p) \mid t \in \mathcal{T}\}. \tag{3}$$

In the following, we describe how to calculate the position $\boldsymbol{f}(t,p)$ of the feet's center of gravity in the image coordinate system. Figure 4(a) shows the parameters used to calculate this position. In the body region image $\boldsymbol{R}(t,p)$ of pedestrian p at time t, our method obtains the image position $\boldsymbol{x}(j,t,p) = (x(j,t,p), y(j,t,p))$ of the point on the feet's contour. Let $\mathcal{J} = \{j\} : j$ be a natural number and $\forall j,k \in \mathcal{J} : j < k \Rightarrow x(j,t,p) < x(k,t,p)$. The origin is the lower left corner of the bounding rectangle of the pedestrian region. Using the

component $y(j,t,p)$, which is the distance from the bottom $(x(j,t,p),0)$ of the bounding rectangle to the feet's contour, we calculate weight $w(j,t,p)$ as

$$w(j,t,p) \sim \mathcal{N}(y(j,t,p)|0,\sigma^2), \tag{4}$$

where $\mathcal{N}()$ is a normal distribution with mean 0 and standard deviation σ. Note that $w(j,t,p)$ satisfies $\sum_{j\in\mathcal{J}} w(j,t,p) = 1$. We obtain the position $\boldsymbol{f}(t,p)$ of the feet's center of gravity in the image coordinate system as follows:

$$\boldsymbol{f}(t,p) = \sum_{j\in\mathcal{J}} w(j,t,p)\boldsymbol{x}(j,t,p). \tag{5}$$

By applying a homography transformation and setting the height on the road surface to 0, we obtain the 3D walking position $\tilde{\boldsymbol{f}}(t,p)$ in the road surface coordinate system.

In the following, we explain why weight $w(j,t,p)$ is assigned to point $\boldsymbol{x}(j,t,p)$ on the feet's contour. Figure 4(b) shows an example when the legs are closed during walking, and (c) shows an example when the legs are open. In the case of closed legs, the candidate contour points mainly appear on the feet, and partially on the hands and other body parts, as shown in the middle part of Fig. 4(b). In the case of open legs, the candidate contour points mainly appear on the feet, and partially on the crotch and other body parts, as shown in the middle part of Fig. 4(c). To suppress the influence of candidate points that do not belong to the feet, we assign small weights to these points in Eq. (5).

The temporal walking position signal $\mathcal{F}(p)$ sometimes contains outliers when the feet's contour is not estimated correctly because of the shadow of a pedestrian on the road surface or markings, such as white lines. Our method detects outliers by applying a Hampel filter and performs a linear interpolation.

2.5 Interaction video sequence generation

In procedure P4, we extract a feature that allows human observers to confirm the body interaction visually. Specifically, we place pedestrians in the same group in a virtual space using a temporal posture signal $\mathcal{V}(p)$ and temporal walking position signal $\mathcal{F}(p)$, and generate an interaction video sequence $\mathcal{S}$ using 3D rendering. In this virtual space, we visualize the temporal posture signal and temporal walking position signal of each pedestrian using the standard body shape parameters, which is the average person's body shape prepared in SMPL, as described in Section 2.3. When we render an interaction video sequence in a virtual space, we always set the virtual camera viewpoint at a fixed position in front of the pedestrian group to capture the physical body interaction which increases the accuracy of conversation activity recognition.

In the following, we explain how to generate an interaction video sequence $\mathcal{S}$. Our method places the 3D vertex $\boldsymbol{v}(t,p) \in \mathcal{V}(t,p) \in \mathcal{V}(p)$ on the body surface obtained in Section 2.3 at the walking position $\tilde{\boldsymbol{f}}(t,p) \in \mathcal{F}(p)$ obtained in Section 2.4. The 3D vertex $\tilde{\boldsymbol{v}}(t,p)$ in the virtual space is converted as follows:

$$\tilde{\boldsymbol{v}}(t,p) = \boldsymbol{v}(t,p) + \tilde{\boldsymbol{f}}(t,p). \tag{6}$$

All vertices $\boldsymbol{v}(t,p)$ in a set $\mathcal{V}(t,p)$ are converted to $\tilde{\boldsymbol{v}}(t,p)$. Suppose that a converted set $\tilde{\mathcal{V}}(t,p)$ consists of $\tilde{\boldsymbol{v}}(t,p)$ and their adjacent vertices. A temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$ in the virtual space is expressed as

$$\tilde{\mathcal{V}}(p) = \{\tilde{\mathcal{V}}(t,p) \mid t \in \mathcal{T}\}. \tag{7}$$

Note that our method determines pedestrian p that belongs to the same group using the distance between the walking positions $\tilde{\boldsymbol{f}}(t,p)$ of pedestrians. After obtaining $\tilde{\mathcal{V}}(p)$ for a pedestrian group, we place each pedestrian that belongs to the same group and perform 3D rendering to generate an image $\boldsymbol{S}(t)$. An interaction video sequence $\mathcal{S}$ for each pedestrian group is expressed as

$$\mathcal{S} = \{\boldsymbol{S}(t) \mid t \in \mathcal{T}\}. \tag{8}$$

The posture parameters are sometimes estimated with an unnaturally large tilt of the human body if a temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$ is directly used for rendering an interactive video sequence. Our method corrects the human body's inclination relative to the road surface by always setting the rotation angle to 0 degrees.

2.6 Conversation activity label classification

In P5, we apply an existing classification network developed in action recognition to determine conversation activity labels using an interaction video sequence $\mathcal{S}$. We use the C3D network [20] that consists of 3D convolution layers designed for action recognition. Our method divides an interaction video sequence into multiple short video sequences, which are input into the C3D network to predict candidate labels that represent conversation activity for each short video sequence. A majority vote among these candidates determines the final label.

In the following, we explain the details of our method for determining the conversation activity label. Our method generates K short video sequences with different initial times from a single interaction video sequence $\mathcal{S}$ during the C3D network training and prediction process. Short video sequence $\hat{\mathcal{S}}$ is expressed as

$$\hat{\mathcal{S}} = \{\boldsymbol{S}(\hat{t}) \mid \hat{t} \in \hat{\mathcal{T}}\}, \tag{9}$$

where $\hat{\mathcal{T}}$ is a set of time points $\hat{t}$ of the image $\boldsymbol{S}(\hat{t})$ that belong to the short video sequence. Our method randomly determines the initial time point $\hat{t}_1$. We generate a short movie sequence $\hat{\mathcal{S}}$ when $\hat{T}(< T)$ images are collected by progressing time at equal intervals I from $\hat{t}_1$. $\hat{T}$ also represents the total number of time points in the short video sequence. During the training process, we train the C3D network using LK short video sequences generated from L interaction video sequences prepared in advance. During the prediction process, we calculate K candidates for the conversation activity label using the input short video sequences generated from an interaction video sequence, and finally determine the output label using majority voting among candidates.

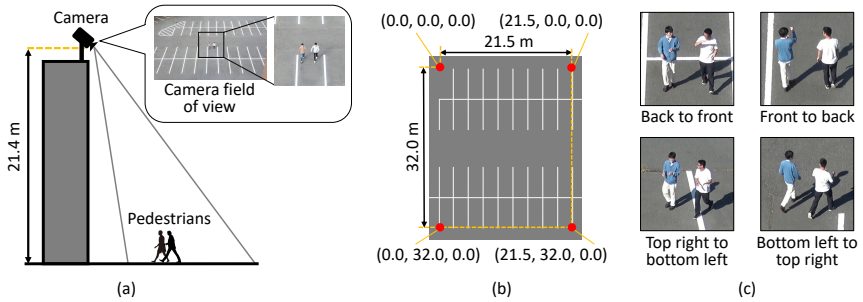

Fig. 5: Camera setting for collecting color video sequences of pedestrian groups while they were walking outdoors and conversing.

3 Experiments

3.1 Dataset

To investigate the effectiveness of our method, we collected color video sequences of pedestrian groups while they were walking outdoors and conversing. Figure 5(a) shows the camera setting. We set the height from the road surface to the camera (SONY, FDR-AX55) to 21.4 m to obtain an overhead view of an outdoor parking lot. The camera resolution was 3840×2160 pixels and the frame rate was 30 fps. Figure 5(b) shows the road surface coordinate system described in Section 2.4. We pre-computed the homography matrix from four white line intersections on the road surface. The camera position in the road surface coordinate system was (10.7, 59.3, 21.4).

We recruited 20 participants (19 men, one woman, 22.6±1.3 years old, university students, Japanese ethnicity). When recruiting the participants, we required that they be somewhat acquainted with each other to avoid a lack of conversation when they first met each other. We controlled the number of pedestrians in a group to a minimum of two participants with whom a conversation could occur. We randomly selected two pedestrians from the 20 participants without duplicates to form a single pedestrian group. We prepared 52 pedestrian groups. We controlled each pedestrian group so that the two participants walked side by side, which is considered to occur most frequently in real scenarios.

We acquired color video sequences of pedestrian groups walking outdoors for each conversation activity label (active conversation, inactive conversation, and no conversation). In one color video sequence, a pedestrian group appeared in the camera's field of view from the start to the end, when it disappeared. To confirm the robustness of the virtual camera viewpoint used in our method, we set four walking paths on the road surface: back to front, front to back, top right to bottom left, and bottom left to top right, as shown in Fig. 5(c). We randomized the order in which the participants walked along each path and the order in which the two participants lined up next to each other. In total, we collected 52 (groups) × 3 (labels) × 4 (walking paths) = 624 color video

Fig. 6: Examples of the pedestrian group video sequences $\mathcal{R}'$ generated from the color video sequences.

sequences. Figure 6 shows examples of the pedestrian group video sequences $\mathcal{R}'$ generated from the collected color video sequences. To generate $\mathcal{R}'$, we set a region of interest for the color video sequence so that two pedestrians that belonged to the same group were within the same field of view using the body region image $\boldsymbol{R}(t, p) \in \mathcal{R}(p)$ estimated in procedure P1.

3.2 Conversation activity labels

When collecting color video sequences, we only instructed the participants on the topic of the conversation and did not give any explanation or instructions regarding the physical body interaction. We set the following conditions for collecting color video sequences for each conversation activity label.

Active conversation:
As a topic of conversation, we instructed the participants to introduce their hobbies while walking. We collected color video sequences while a pedestrian talked about a hobby, the other pedestrian responded to it, and started a new conversation about a hobby.

Inactive conversation:
As a topic of conversation, we instructed the participants to talk about topics of little interest to each other while walking. The topic was chosen by the participants from several candidate topics prepared in advance (e.g., economic situation and political situation in a country that the participants had never visited and had almost no knowledge of).

No conversation:
We instructed the participants not to engage in any conversation while walking.

Fig. 7: Examples of interaction video sequences $\mathcal{S}$.

We randomized which pedestrians in the group initiated the conversation when collecting active and inactive labels.

3.3 Experimental conditions

In the following, we describe the experimental conditions for procedures P1 through P3. We used the default parameters provided for PHALP in P1 and P2. The window size of the Hampel filter in P2 and P3 was 5. We set the body shape parameters of SMPL to the default parameters provided by PHALP. We automatically determined the σ of Eq. (4) in P3 according to the height of the pedestrian's bounding rectangle. Specifically, σ increased as the height increased and σ decreased as the height decreased.

Next, we describe how to determine the virtual camera viewpoint for generating interaction video sequences in procedure P4. We determined the direction in which a pedestrian group walks on a road surface by fitting a straight line using the group's center positions at all time points. We always kept the virtual camera viewpoint at a distance of 4.25 m from the center position in the direction of the pedestrian group. The height of the virtual camera viewpoint was 0.85 m from the road surface. Figure 7 shows examples of the interaction video sequences $\mathcal{S}$. The color scheme for each pedestrian was either light red or light blue and was determined randomly without duplication. We believe that human observers can visually confirm the posture among pedestrians, such as arm bending and face orientation, and the positioning of the pedestrians in each group, from the interaction video sequences in the figure.

The C3D network [20] in procedure P5 consisted of four convolution layers, four pooling layers, and two affine layers. The filter size for 3D convolution was

$3 \times 3 \times 3$. Time length $\hat{T}$ of a short video sequence $\hat{\mathcal{S}}$ was 16. We set the array size of the short video sequences to 100 (pixels) $\times$ 100 (pixels) $\times$ 3 (colors) $\times$ 16 (time points). We set $I = 18$ and $K = 50$ for the parameters described in Section 2.6. We used RMSprop as the optimizer when training the C3D network, with a learning rate of 0.0001 and mini-batch size of 16. We trained the C3D network from scratch.

We applied leave-one-group-out when evaluating the accuracy of conversation activity recognition. Specifically, we used 12 interaction video sequences generated from one pedestrian group for the prediction process and $L = 612$ interaction video sequences generated from the remaining 51 pedestrian groups for the training process. We repeated the training and prediction processes for all 52 pedestrian groups. We prepared 3 (labels) $\times$ 4 (walking paths) = 12 interaction video sequences per pedestrian group.

We evaluated the computational cost of our method on a PC equipped with a GPU (RTX 2080 Ti) and CPU (i9-9940X). The processing time was 0.29 seconds for P1, 0.66 seconds for P2, 0.05 seconds for P3, and 0.58 seconds for P4 per video sequence frame. The processing time for P5 was 0.01 seconds per short video sequence during prediction. The total GPU memory usage was 4.7 GB.

3.4 Basic performance

We evaluated the effectiveness of our method using interaction video sequences as features. For comparison, we used the following features to calculate the accuracy of conversation activity recognition.

M1: Interaction video sequence $\mathcal{S}$
We used $\mathcal{S}$ generated in procedure P4 of our method as the feature. Specifically, we generated short video sequences $\hat{\mathcal{S}}$ in procedure P5 from $\mathcal{S}$. The array size of the short video sequence was 100 (pixels) $\times$ 100 (pixels) $\times$ 3 (colors) $\times$ 16 (time points).

M2: Pedestrian group video sequence $\mathcal{R}'$
We used $\mathcal{R}'$, which represents the appearance of the pedestrian group, as the feature. Examples of $\mathcal{R}'$ were already shown in Fig. 6. We directly passed the pedestrian group video $\mathcal{R}'$ to procedure P5 and generated short video sequences from $\mathcal{R}'$. The array size of the short video sequence was 100 (pixels) $\times$ 100 (pixels) $\times$ 3 (colors) $\times$ 16 (time points).

M3: Temporal posture signal $\mathcal{V}(p)$
We used $\mathcal{V}(p)$ estimated from each pedestrian that belonged to the same group as the feature. Specifically, we directly passed $\mathcal{V}(p)$ estimated in procedure P2 to procedure P5 and then generated short temporal signals. The array size of the short temporal signal was 6890 (vertices) $\times$ 2 (pedestrians) $\times$ 3 (components) $\times$ 16 (time points).

M4: Temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$
We used $\tilde{\mathcal{V}}(p)$, combining a temporal posture signal $\mathcal{V}(p)$ with the temporal walking position signal $\mathcal{F}(p)$ estimated from each pedestrian that belonged to the same group as the feature. Specifically, we directly passed $\tilde{\mathcal{V}}(p)$ generated in procedure P4 to procedure P5 and then generated short temporal

signals. The array size of the short temporal signal was 6890 (vertices) × 2 (pedestrians) × 3 (components) × 16 (time points).

We input each feature into the C3D network to predict the conversation activity label in P5. We calculated accuracy using the number of correctly predicted conversation activity labels. Because there was random sampling when we extracted each feature, we set the number of trials used to calculate recognition accuracy to 10. In M3 and M4, to align the dimensionality with other features, we randomly sampled 5000 vertices and then transformed the array size from 5000 × 2 × 3 × 16 to 100 × 100 × 3 × 16. In each accuracy evaluation trial, we assumed that the vertices sampled in all short temporal signals were the same. The other experimental conditions were the same as those described in Section 3.3.

Table 1 shows the accuracy of using each feature in conversation activity recognition. Recognition accuracy was 76.2±0.7% for interaction video sequence $\mathcal{S}$ of M1, 57.3±1.3% for pedestrian group video sequence $\mathcal{R}'$ of M2, 72.9±0.9% for temporal posture signal $\mathcal{V}(p)$ of M3, and 74.1±0.7% for temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$ of M4. In all cases, we confirmed that our method M1 was more accurate than M2, M3, and M4. These results indicate that using a feature of an interaction video sequence generated by our method was more effective in recognizing conversation activity than using a feature of a pedestrian group video sequence, a temporal posture signal, or a temporal posture and walking position signal.

Instead of C3D, we applied TimeSformer [1] as a video action recognition method and LSTM [8] as a time series analysis method. TimeSformer performed fine-tuning on a model pre-trained with Kinetics-400, whereas LSTM trained a model from scratch. The recognition accuracies were 71.8±0.7% for TimeSformer and 67.1 ± 1.1% for LSTM. Our method obtained higher recognition accuracy (76.2±0.7%) than the existing methods. The GPU memory usage was 1.3 GB for C3D used in our method, 6.1 GB for TimeSformer, and 0.7 GB for LSTM. We believe that our method is reasonable in terms of the trade-off between accuracy and memory usage.

We evaluated the recognition accuracy of our method for the case of several groups walking simultaneously. The number of groups in each frame ranged from 0 to 3. We used a total of 120 groups. The accuracy of our method was 67.4±0.2%. Although our method performed well in this case with minimal occlusion, it is important to note that real-world scenarios often involve heavy occlusion caused by people overlapping. This presents a significant limitation that we need to address in future work. For practical applications, we must develop methods for various scenarios, such as heavy occlusion and interaction with objects such as shopping trolleys.

3.5 Evaluation of different virtual camera viewpoints

We evaluated the accuracy of conversation activity recognition for different virtual camera viewpoints when generating an interaction video sequence in procedure P4. We set the positions of the virtual camera viewpoints on C1 front,

Table 1: Comparison of the accuracy of conversation activity recognition using each feature.

Feature for conversation activity recognition	Accuracy (%)
M1: Interaction video sequence $\mathcal{S}$	**76.2±0.7**
M2: Pedestrian group video sequence $\mathcal{R}'$	57.3±1.3
M3: Temporal posture signal $\mathcal{V}(p)$	72.9±0.9
M4: Temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$	74.1±0.7

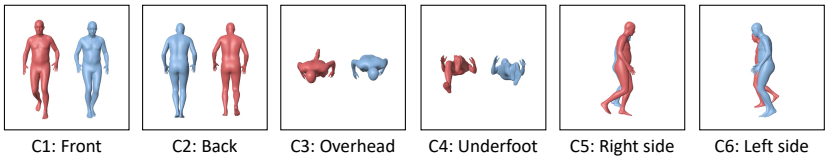

Fig. 8: Examples of interaction video sequences generated from different virtual camera viewpoints in procedure P4 of our method.

C2 back, C3 overhead, C4 underfoot, C5 right side, and C6 left side. Figure 8 shows interaction video sequences generated using these virtual camera viewpoints. We changed only the position of the virtual camera viewpoint; the other experimental conditions were the same as those described in Section 3.4.

Table 2 shows the accuracy for each virtual camera viewpoint when generating interaction video sequences. We confirmed that C1, in which the virtual camera viewpoint was the front of the pedestrian group, had higher recognition accuracy than C2, C3, C4, C5, and C6, in which the virtual camera viewpoint was not the front of the pedestrian group. Furthermore, we checked the recognition accuracy of our method C1 for each walking path in Fig. 5(c). We achieved the same level of accuracy for all walking paths. Based on these results, when generating interaction video sequences in procedure P4, placing the virtual camera viewpoint in a position that always captured a pedestrian group from the front led effectively to the recognition of conversation activity.

4 Conclusions

We proposed a method for recognizing conversation activity in a group of pedestrians walking outdoors using interaction video sequences that represent human body interactions. The experimental results demonstrated that our method is superior to the alternative techniques using pedestrian body region video sequences or temporal posture and walking position signals in conversation activity recognition. We believe that our method can be implemented in a variety of potential applications in addition to the marketing applications described in Section 1. For

Table 2: Accuracy of conversation activity recognition when generating interaction video sequences from different virtual camera viewpoints in P4.

Virtual camera viewpoint	Accuracy (%)
C1: Front	**76.2±0.7**
C2: Back	74.8±0.3
C3: Overhead	70.2±0.8
C4: Underfoot	72.0±0.9
C5: Right side	40.0±1.5
C6: Left side	48.2±2.6

example, we considered medical applications for dementia checking, office applications for mental health checking, and educational applications for bullying detection. In future work, we intend to develop a method to recognize conversation activity at multiple levels and a robust method for occlusion. We will expand evaluations by increasing the number of pedestrians in the same group and changing the positional relationship of pedestrians within a group. We will perform a performance comparison with group activity recognition methods, for example, ARG [21], Actor-Transformers [4], GroupFormer [10], DIN [22], and KRGFormer [14]. We appreciate Professor Yoshio Iwai's valuable advice and suggestions during this study. We would like to thank Mr. Norihiko Torii, Mr. Tomohiro Miyake, and Mr. Osamu Yoshimura of SEIRYO ELECTRIC Corporation for their helpful advice on this paper.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? Proceedings of the International Conference on Machine Learning (2021)
2. Chamveha, I., Sugano, Y., Sato, Y., Sugimoto, A.: Social group discovery from surveillance videos: A data-driven approach with attention-based cues. In: Proceedings of the British Machine Vision Conference. pp. 1–12 (2013)
3. Ehsanpour, M., Saleh, F., Savarese, S., Reid, I., Rezatofighi, H.: JRDB-Act: A large-scale dataset for spatio-temporal action, social group and activity detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 20951–20960 (2022)
4. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.M.: Actor-transformers for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 836–845 (2020)
5. Ge, W., Collins, R.T., Ruback, R.B.: Vision-based analysis of small groups in pedestrian crowds. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(5), 1003–1016 (2012)
6. Han, R., Yan, H., Li, J., Wang, S., Feng, W., Wang, S.: Panoramic human activity recognition. In: Proceedings of the European Conference on Computer Vision. pp. 224–261 (2022)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)

8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
9. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: Advances in Neural Information Processing Systems. vol. 1, p. 1216–1224 (2010)
10. Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: Groupformer: Group activity recognition with clustered spatial-temporal transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13648–13657 (2021)
11. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 1–16 (2015)
12. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992)
13. Odashima, S., Shimosaka, M., Kaneko, T., Fukui, R., Sato, T.: Collective activity localization with contextual spatial pyramid. In: Proceedings of the European Conference on Computer Vision. pp. 243–252 (2012)
14. Pei, D., Huang, D., Kong, L., Wang, Y.: Key role guided transformer for group activity recognition. IEEE Transactions on Circuits and Systems for Video Technology **33**(12), 7803–7818 (2023)
15. Qing, L., Li, L., Xu, S., Huang, Y., Liu, M., Jin, R., Liu, B., Niu, T., Wen, H., Wang, Y., Jiang, X., Peng, Y.: Public life in public space (PLPS): A multi-task, multi-group video dataset for public life research. In: Proceedings of the International Conference on Computer Vision Workshops. pp. 3611–3620 (2021)
16. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3D appearance, location and pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2740–2749 (2022)
17. Rota, P., Conci, N., Sebe, N.: Real time detection of social interactions in surveillance video. In: Proceedings of the European Conference on Computer Vision. pp. 111–120 (2012)
18. Solera, F., Calderara, S., Cucchiara, R.: Socially constrained structural learning for groups detection in crowd. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(5), 995–1008 (2016)
19. Su, J., Huang, J., Qing, L., He, X., Chen, H.: A new approach for social group detection based on spatio-temporal interpersonal distance measurement. Heliyon **8**(10), e11038 (2022)
20. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497 (2015)
21. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9956–9966 (2019)
22. Yuan, H., Ni, D., Wang, M.: Spatio-temporal dynamic inference network for group activity recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7456–7465 (2021)
23. Zanlungo, F., Brščić, D., Kanda, T.: Pedestrian group behaviour analysis under different density conditions. Transportation Research Procedia **2**, 149–158 (2014)
24. Zanotto, M., Bazzani, L., Cristani, M., Murino, V.: Online bayesian nonparametrics for group detection. In: Proceedings of the British Machine Vision Conference. pp. 1–12 (2012)