Reducing Computational Cost in Pedestrian Conversation Activity Recognition through Skeleton Spatiotemporal Graphs

Tsubasa Kondo*, Michiko Inoue*, Shunsuke Yoneda† and Masashi Nishiyama*
*Graduate School of Sustainability Science, Tottori University, Japan
†Organization for Information Strategy and Management, Tottori University, Japan

***Abstract*—The existing method for recognizing conversation activity in walking pedestrian groups uses heavy three-dimensional mesh reconstruction and rendering, which significantly increases the computational cost. We investigate a lightweight framework that replaces skinned multi-person linear model mesh structures using skeleton spatiotemporal graphs extracted from general color video sequences to recognize the conversation activity labels. Compared with the existing method, our approach reduces inference time and graphics processing unit memory while maintaining equal or higher accuracy, which makes low computational cost conversation activity recognition feasible.**

***Index Terms*—Recognition, Conversation activity, Computational cost, Skeleton, Spatiotemporal graphs**

I. Introduction

A growing demand exists for a technique that automatically recognizes conversation activity inside pedestrian groups walking outdoors. Such a technique acquires color video sequences using a general camera from a standoff distance and extracts visual features that reflect the group's conversation activity. To the best of our knowledge, only one existing method [1] has addressed conversation activity recognition for walking groups using video sequences. This method uses a succession of whole-body movements, termed body interaction, as the visual features based on McNeill's [2] finding that gestures are bodily movements that accompany speech and are helpful for the analysis of conversation. The existing method extracts the body gestures of pedestrians using a three-dimensional (3D) human body model and inputs the resulting interaction video sequence as the visual features. In addition to its high accuracy, this design enables developers to visually confirm which body parts in the 3D model contribute to conversation activity recognition.

However, the existing method [1] was not designed for scenarios with a limited computation time and graphics processing unit (GPU) memory usage. We identify two dominant bottlenecks in the processing procedures of the existing method. First, to produce a 3D human model for every pedestrian in the video sequence, the method estimates the skinned multi-person linear model (SMPL) pose-and-shape parameters [3] and converts them into a mesh structure. Although low-dimensional SMPL parameters can be inferred stably and accurately, the computation required is high. Second, when the visual features are built, the existing method renders each mesh structure onto image planes to form an interaction video sequence; the rendering cost increases as the video-sequence length increases, and thus demands extra computation and GPU memory.

To overcome such resource-limited settings, we investigate a technique, referred to hereafter as our method, that replaces the SMPL mesh pipeline with a much lighter skeleton representation. Specifically, we encode the pedestrian group's body gesture interaction as a skeleton spatiotemporal graph built from body-joint keypoints and evaluate its effectiveness for conversation activity recognition. We verify the following two points:

- **I1**: In the body movement parameter inference stage, our method lowers computation time and GPU memory usage compared with the existing method [1].
- **I2**: In the feature extraction stage, our method lowers computation time and GPU memory usage compared with the existing method [1].

We also confirm that our method achieves recognition accuracy on a par with, or superior to, the existing method.

II. Our method

A. Overview

We design a new approach to achieve points I1 and I2 described in Section I while preserving the recognition accuracy of the existing method [1]. Specifically, we design a method that represents the body interaction effective for conversation activity recognition as a skeleton spatiotemporal graph and uses this representation as the visual feature.

Figure 1 shows an overview of our method. We use a skeleton spatiotemporal graph as the feature for conversation activity recognition. The processing procedure is as follows: First, Procedure P1 detects the bounding box of each pedestrian in the group using YOLOX-X [4]. Next, Procedure P2 extracts the skeleton of each pedestrian using ViTpose [5] at each time point and Procedure P3 connects these skeletons along the temporal direction to generate a spatiotemporal graph. Finally, Procedure P4 introduces the resulting spatiotemporal graph to predict conversation activity using ST-GCN [6].

B. Skeleton spatiotemporal graph

We aim to achieve points I1 and I2 described in Section I using a skeleton spatiotemporal graph. Specifically, at each time point, we represent the skeleton as the image-plane

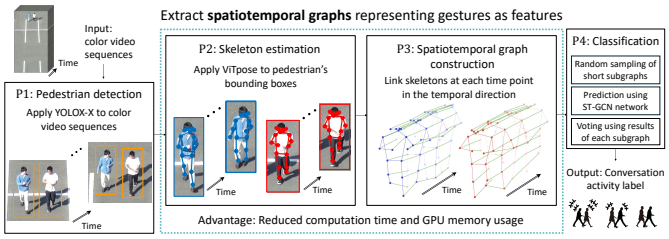

Fig. 1. Overview of our method

positions of keypoints, such as the center of mass of the head and the body joints, together with connectivity among these keypoints. The body can be represented by the skeleton at each time point without the computationally intensive, high-precision estimation of SMPL pose-and-shape parameters, and therefore we expect to reduce the computation time required to estimate the bodies of pedestrians in a video sequence.

Our method suppresses the large amount of GPU memory usage that would otherwise be required to process the body-interaction video sequence rendered from the SMPL model by linking skeletons along the temporal axis to form a skeleton spatiotemporal graph. Specifically, within the same frame of the video sequence, we connect the keypoints of each skeleton in the spatial domain, and between temporally adjacent frames, we connect the corresponding keypoints in the temporal domain, thereby constructing the graph. The skeleton spatiotemporal graph expresses body interaction without performing 3D rendering, and thus this graph reduces both computation time and GPU memory usage.

C. Pipeline of our method

1) P1: Pedestrian detection: We detect a bounding box that contains the entire body region of each pedestrian from the camera video sequence. As the detector, we use YOLOX-X [4], which estimates whole-body bounding boxes. For a pedestrian p, we define the resulting video sequence of estimated detected bounding boxes as

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

where $\boldsymbol{R}(t,p)$ is the region image of pedestrian p at time point t and $\mathcal{T}$ is the set of time points at which the region images were acquired. The total number of elements in $\mathcal{T}$ is denoted by T. This T also corresponds to the duration during which pedestrian p remains in view, from the moment the person appears until the individual leaves the frame in the video sequence. Our method performs pedestrian tracking concurrently with detection to associate the same individual across frames. We adopt ByteTrack [7] as the tracking algorithm.

2) P2: Skeleton estimation: We estimate the skeleton of pedestrian p from the bounding-box video sequence $\mathcal{R}(p)$. We adopt ViTpose [5] as the skeleton-estimation method. For the region image $\boldsymbol{R}(t,p) \in \mathcal{R}(p)$ at time point t, we estimate the image-plane locations of the pedestrian's keypoints. The skeleton is represented by a spatial graph $\mathcal{S}(t,p)$ that contains these keypoint positions. This spatial graph connects the keypoints of the skeleton in the spatial domain within the same time point of the video sequence.

3) P3: Spatiotemporal graph construction: We generate a spatiotemporal graph for pedestrian p from the time-series signal of skeletons $\mathcal{S}(t,p)$. This spatiotemporal graph links the corresponding keypoints between adjacent time points along the temporal domain. We express the spatiotemporal graph $\mathcal{G}(p)$ for pedestrian p as

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

4) P4: Classification: We input the spatiotemporal graph $\mathcal{G}(p)$ and output a label that represents conversation activity. As the classification network, we use ST-GCN [6], which simultaneously performs spatial convolution and temporal convolution. Specifically, we randomly generate multiple short subgraphs from the spatiotemporal graph, feed each subgraph into ST-GCN, and obtain multiple label candidates representing conversation activity. Finally, we use a majority vote over these candidates to produce the conversation activity label. The output label is one of the three categories: active, inactive, or no conversation.

During both the training and inference of ST-GCN, we randomly generate K short subgraphs $\hat{\mathcal{G}}(p)$ that have different initial time points from a single spatiotemporal graph $\mathcal{G}(p)$. We define a short subgraph $\hat{\mathcal{G}}(p)$ as

$$\hat{\mathcal{G}}(p) = \mathrm{randamsampling}(\mathcal{G}(\mathrm{p})) = \{\mathcal{S}(\hat{\mathrm{t}},\mathrm{p}) \mid \hat{\mathrm{t}} \in \mathcal{T}\}, \tag{3}$$

where $\hat{\mathcal{T}}$ is the set of time points $\hat{t}$ that belong to a short subgraph. One short subgraph $\hat{\mathcal{G}}(p)$ is generated by the random selection of an initial time point followed by the advancement of the time point at fixed interval I until the collection of $\hat{T}$ spatial skelton graphs $\mathcal{S}(\hat{t},p)$. As noted in Section II-C1, the total number of time points in the set $\mathcal{T}$ is T; consequently, the spatiotemporal graph $\mathcal{G}(p)$ also contains T time points. Let $\hat{T}$ be the number of time points contained in one short subgraph, where $\hat{T} < T$. During the training phase, our method generates K short subgraphs $\hat{\mathcal{G}}(p)$ with different initial time points from a single training spatiotemporal $\mathcal{G}(p)$, and trains ST-GCN using LK short subgraphs generated from L training spatiotemporal graphs. During the inference phase, we generate K short subgraphs from an input spatiotemporal graph, obtain K label candidates for conversation activity, and finally output the label using majority voting over these candidates.

III. Existing Method

We describe the procedures of the existing method [1]. Figure 2 shows an overview of the existing method. We describe each procedure below.

- $\bar{\text{P}}1$ *Pedestrian segmentation*: Given a color video sequence, this procedure estimates human body regions by segmenting the whole body of each pedestrian that appears in the video sequence using Mask R-CNN [8].
- $\bar{\text{P}}2$ *SMPL estimation*: From the appearance of the pedestrian regions obtained in $\bar{\text{P}}1$, this procedure estimates the SMPL model parameters that describe body pose

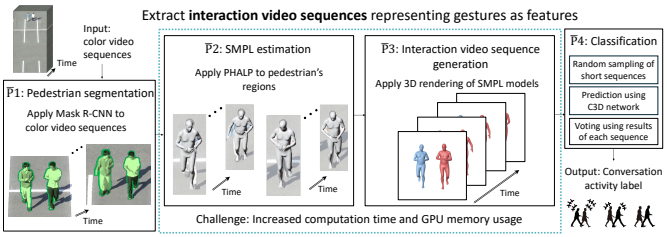

Fig. 2. Overview of the existing method

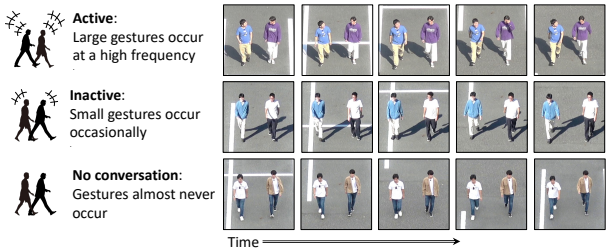

Fig. 3. Examples of the change of appearance of pedestrian groups.

and shape using PHALP [9] and then converts these parameters into a 3D mesh structure.

- P̄3 *Interaction video sequence generation*: Using the 3D mesh structure converted from the SMPL model in P̄2, this procedure renders the mesh structure onto the image plane to create the interaction video sequence. The existing method refers to the video sequence that represents body interaction as the interaction video sequence.
- P̄4 *Classification*: The interaction video sequence generated in P̄3 is fed into the classification network of C3D [10], which outputs a conversation activity label. First, multiple short video sequences generated from a single interaction video sequence are input into the network. Second, a label is produced for each short video sequence and the final conversation activity label is determined by majority vote.

IV. Experiments

A. Dataset

To verify the effectiveness of our method, we used the dataset collected in experiments for the existing method [1]. Figure 3 shows sample images in which the regions of pedestrian groups are highlighted. We prepared three conversation activity labels: active, inactive, and no conversation. The *active* label indicates that the pedestrian group is engaged in a lively conversation on topics of mutual interest. The *inactive* label indicates that the group is not engaged in a lively conversation, for example, because the topic is of little interest. The *no conversation* label indicates that no conversation is taking place. The dataset comprised $52 \times 4 \times 3 = 624$ color video sequences collected from 52 pedestrian pairs and the total duration of all sequences was approximately 3.4 hours.

B. Experimental conditions

In the experiments, we used the following procedures of our method and the existing method to confirm whether points I1 and I2 described in Section I were achieved.

- Evaluation of I1: comparison between Procedure P2 of our method and Procedure P̄2 of the existing method in the body movement parameter inference stage.
- Evaluation of I2: comparison between Procedure P3 of our method and Procedure P̄3 of the existing method in the feature extraction stage.

We also compared the total processing times of all procedures for both methods.

We evaluated recognition accuracy using the leave-one-pedestrian-pair-out scheme for both our method and the existing method. Moreover, we repeated the recognition accuracy evaluation ten times because random sampling involved feature construction for both methods.

C. Parameters of our method

The ST-GCN [6] network model was composed of ten spatiotemporal convolution layers, one average-pooling layer, and one fully connected layer. A residual block was inserted after each spatiotemporal graph convolution. In the spatiotemporal convolution layers, the kernel size was set to 3 for spatial convolution and 9 for temporal convolution. For the short subgraph parameters described in Section II-C4, the settings were as follows: $I = 18$, $K = 50$, and $\hat{T} = 16$. The resulting input short time-series array had the shape of 16 (time points) $\times$ 17 (keypoints) $\times$ 2 (x, y channels). SGD was used as the optimizer with the following hyperparameters: learning rate 0.00002, momentum 0.9, weight decay 0.0004, and mini-batch size 32. For the existing method, the default hyperparameter settings were those reported in [1].

D. Results

Table I presents the computation times of our method during inference, whereas Table II presents the corresponding times for the existing method. The experimental result showed that Procedure P2 of our method required less computation time than Procedure P̄2 of the existing method. The experimental result showed that Procedure P3 of our method was much faster than Procedure P̄3 of the existing method. When the computation times were summed, our method was approximately 10.5 times faster than the existing method.

Table III lists the GPU memory usage of our method during inference and Table IV lists that of the existing method. Procedure P2 of our method consumed less GPU memory than Procedure P̄2 of the existing method. Procedure P3 of our method also consumed less GPU memory than Procedure P̄3 of the existing method. A comparison of total GPU memory usage indicated that our method reduced the requirement to approximately 61 % of that of the existing method.

Table V compares conversation activity recognition accuracies between our method and the existing method. The results confirm that our method's recognition accuracy was comparable to or higher than that of the existing method.

TABLE I
COMPUTATION TIME OF OUR METHOD DURING INFERENCE

	Procedure	Time	Unit
P1	Pedestrian detection	0.085	seconds per frame
P2	Skeleton estimation	0.021	seconds per frame
P3	Spatiotemporal graph construction	0.000	seconds per frame
P4	Classification	0.001	seconds per short subgraph
Total		0.107	seconds

TABLE II
COMPUTATION TIME OF THE EXISTING METHOD DURING INFERENCE

	Procedure	Time	Unit
$\bar{\text{P}}1$	Pedestrian segmentation	0.249	seconds per frame
$\bar{\text{P}}2$	SMPL estimation	0.326	seconds per frame
$\bar{\text{P}}3$	Interaction video sequence generation	0.542	seconds per frame
$\bar{\text{P}}4$	Classification	0.007	seconds per short video sequence
Total		1.124	seconds

TABLE III
GPU MEMORY USAGE OF OUR METHOD DURING INFERENCE

	Procedure	Memory usage (MiB)
P1	Pedestrian detection	978
P2	Skeleton estimation	1002
P3	Spatiotemporal graph construction	0
P4	Classification	1194
Total		3174

TABLE IV
GPU MEMORY USAGE OF THE EXISTING METHOD DURING INFERENCE

	Procedure	Memory usage (MiB)
$\bar{\text{P}}1$	Pedestrian segmentation	1020
$\bar{\text{P}}2$	SMPL estimation	1220
$\bar{\text{P}}3$	Interaction video sequence generation	1227
$\bar{\text{P}}4$	Classification	1690
Total		5157

V. CONCLUSIONS

We investigated the effectiveness of a technique that recognizes conversation activity using skeleton spatiotemporal graphs, estimated from color video sequences, as informative and compact features. The experimental results confirmed that our method significantly reduced computation time and GPU memory usage while achieving recognition accuracy comparable to or higher than the existing method.

In future work, we intend to apply conversation activity recognition to various video sequences collected in more practical scenarios. We will extend the framework to pedestrian groups with three or more members and scenes in which group membership changes over time. We will also investigate adaptive spatiotemporal graph construction that handles occlusion, missing keypoints, and large pose variations.

ACKNOWLEDGMENT

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.

TABLE V
ACCURACIES OF OUR METHOD AND THE EXISTING METHOD

Method	Accuracy (%)
Our method	80.3 ± 0.8
Existing method [1]	76.2 ± 0.7

REFERENCES

[1] W. Ganaha, T. Ozaki, M. Inoue, and M. Nishiyama, "Conversation activity recognition using interaction video sequences in pedestrian groups," in *Proceedings of the 27th International Conference on Pattern Recognition, Part XIV*, 2024, pp. 359–374.

[2] D. McNeill, *Hand and mind: What gestures reveal about thought*, University of Chicago Press, 1992.

[3] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 1–16, 2015.

[4] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[5] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 38571–38584.

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018, vol. 32, pp. 7444–7452.

[7] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object tracking by associating every detection box," in *Proceedings of the European Conference on Computer Vision*, 2022, vol. 13682, pp. 1–21.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[9] J. Rajasegaran, G. Pavlakos, A. Kanazawa, and J. Malik, "Tracking people by predicting 3D appearance, location and pose," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2740–2749.

[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.