Visualizing Body-Interaction Importance for Conversation Activity Recognition in Pedestrian Groups

Wataru Ganaha*, Takumi Ozaki*, Masashi Nishiyama*
*Graduate School of Engineering, Tottori University, Japan

***Abstract*—Recognizing conversation activity within walking pedestrian groups is crucial for understanding human interactions. While existing methods achieve high accuracy using physical body interactions, it remains unclear which body parts and temporal states contribute most to the classification. In this paper, we propose a visualization method based on Integrated Gradients (IG) to analyze body-interaction importance in video sequences. To resolve the spatial scattering and temporal discontinuity caused by directly applying IG, we introduce spatial enhancement using kernel density estimation and a temporal integration of importance maps. Our method visualizes pixels with large attribution magnitudes, suggesting that the network may be sensitive to specific physical states, such as hand gestures and head orientations, for different activity labels. This approach provides insights into potentially effective features for recognition and the classification tendencies of the network.**

***Index Terms*—Conversation activity recognition, Pedestrian groups, Visualization, Integrated Gradients, Body interaction**

I. Introduction

Video analysis can automatically recognize interactions within groups walking outdoors and support a deeper understanding of human behavior. We call a video sequence of a walking pedestrian group in which conversation may occur a pedestrian group video sequence. This study focuses on conversation activity among the various interactions in such sequences. Conversation activity indicates whether conversation occurs and, if so, whether it is active or inactive.

We previously hypothesized that temporal changes in gestures and mutual body orientation among pedestrians provide cues for conversation activity recognition [1]. During active conversation, participants may raise their hands or elbows while gesturing and turn their heads or upper bodies toward one another more often than during inactive conversation. We use physical body interaction to denote these joint temporal patterns.

To test this hypothesis, our previous method recognizes conversation activity using only physical body interaction [1]. It creates an interaction video sequence in a virtual space from a pedestrian group video sequence, samples frames at regular intervals to form short video sequences, and sends them to a C3D model. The model assigns active conversation, inactive conversation, or no conversation.

Although our previous evaluation demonstrated high recognition accuracy [1], prior work has neither visualized the body-interaction importance of every pixel at each time point nor examined body states within high-importance regions.

This paper therefore proposes a visualization method based on Integrated Gradients (IG) [2], which attributes a model output to input features through gradients. Direct application of IG causes spatial scattering of low-importance pixels and temporal discontinuity between short video sequences. The proposed method addresses these problems through spatial enhancement and temporal integration. We then analyze which body parts and states exhibit large attribution magnitudes.

II. Conversation Activity among Pedestrians

A. Background of Physical Body Interaction

This section examines the two components of physical body interaction defined in Section I. Gestures are body movements that accompany speech, and research has established their value in conversation analysis [3]. We therefore expect gestures expressed through hand and elbow movements to provide cues for analyzing conversation activity. Mutual body orientation also reflects interpersonal relations within a group. Previous studies have demonstrated the usefulness of body orientation for pedestrian group behavior analysis [4] and group detection in surveillance video [5]; we expect it to benefit conversation activity recognition as well.

During an active conversation, speech-related gestures may raise the hands or elbows, while the head and upper body may turn toward the conversation partner more often than during an inactive conversation. In contrast, a group without conversation may show no gestures, and the heads may remain oriented in the walking direction. A representation that combines gestures with mutual body orientation can therefore help distinguish both the presence and activity level of a conversation.

B. Recording Pedestrian Group Video Sequences

We controlled the conversation topic to define the conversation activity labels for the pedestrian group video sequences used in the visualization experiments. We applied the following recording conditions.

Active conversation: We asked each pair to introduce their hobbies to one another while walking. Either participant could begin the conversation.

Inactive conversation: We asked each pair to discuss a topic of little interest to either participant. The pair selected a topic from a prepared list, such as the economic or political situation in a country that neither participant had visited.

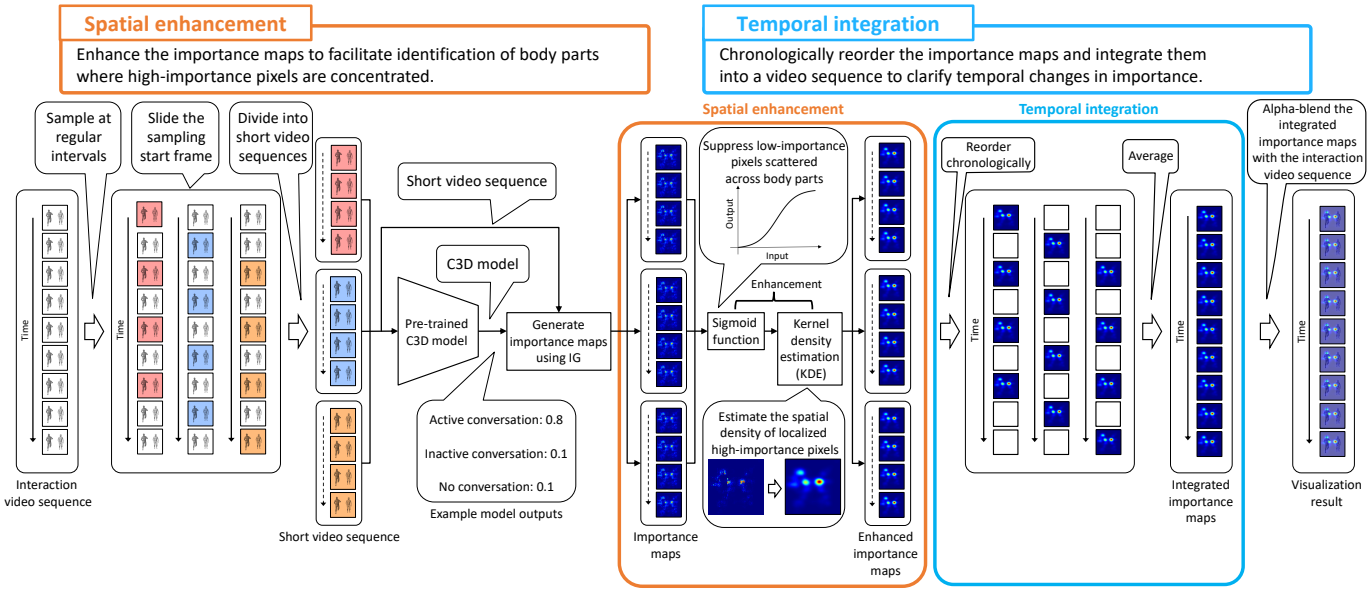

Fig. 1. Overview of the proposed method for visualizing body-interaction importance.

No conversation: We asked each pair to walk without speaking.

Apart from the condition-specific instructions described above, we neither explained physical body interaction nor asked the participants to perform predefined body movements. Thus, the recording condition, rather than a subjective judgment of visible behavior by an external observer, determines each conversation activity label.

III. Proposed Method

A. Preparation

First, we generate an interaction video sequence from each pedestrian group video sequence recorded under one of the recording conditions in Section II-B. The interaction video sequence represents physical body interaction within the group. Starting at a given frame, we then sample F frames at intervals of I frames to generate a short video sequence. Each short video sequence has an array size of h pixels $\times$ w pixels $\times$ F frames $\times$ 3 RGB channels.

We feed each short video sequence to the recognition model developed in our previous study [1], which was initialized from a pre-trained C3D model [6]. The model returns a model output vector for active conversation, inactive conversation, and no conversation. We trained and evaluated the C3D model using leave-one-pair-out cross-validation. In each fold, all video sequences from one pair formed the evaluation set, while video sequences from the remaining pairs formed the training set; we repeated this procedure with each pair as the held-out pair.

B. Visualizing Body-Interaction Importance

1) Visualization Challenges: We apply IG to each short video sequence to calculate the importance of every pixel at every time point and generate heatmap-style importance maps. We call the alpha-blended combination of a short video sequence and its importance maps a visualization result. Direct application of IG presents two challenges.

Spatial challenge: IG scatters low-importance pixels across several body parts and intersperses them with high-importance pixels. This scattering makes it difficult to determine which body parts contain high-importance pixels.

Temporal challenge: Each short video sequence produces an independent set of importance maps. Without aligning the maps to their source frames, we cannot follow continuous changes in body-interaction importance throughout the complete interaction video sequence.

2) Overview: Figure 1 illustrates the proposed pipeline. First, we sample an interaction video sequence at regular intervals while sliding the sampling start frame to generate multiple short video sequences. The C3D model produces a model output vector for each sequence, and the IG-based procedure generates the corresponding importance maps.

IG yields an attribution tensor with the same dimensions as the input: $h \times w \times F \times 3$. Because the attributions can take negative values, we use their absolute values and sum them along the RGB-channel dimension to obtain one two-dimensional importance map for each frame. Therefore, the resulting maps represent attribution magnitude and do not distinguish positive from negative effects on the model output for the predicted class. Hereafter, we refer to this attribution magnitude as body-interaction importance. For spatial enhancement, we sequentially apply a sigmoid function and kernel density estimation (KDE) to suppress low-importance pixels scattered across body parts. For temporal integration, we reorder the maps chronologically and average maps that correspond to the same source frame. Finally, we alpha-blend

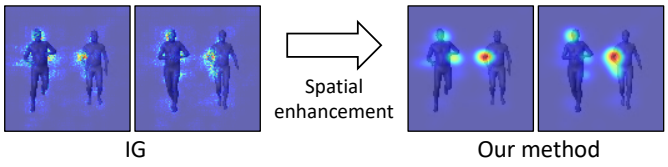

Fig. 2. Effect of spatial enhancement on importance maps.

the integrated importance maps with the interaction video sequence.

3) Spatial Enhancement: Spatial enhancement clarifies regions in which high-importance pixels are concentrated. First, we normalize every pixel value in an importance map to the range from 0 to 1. We then apply the sigmoid function shown in Eq. (1).

$$y = \frac{1}{1 + e^{-a(x-0.5)}} \quad (a > 0). \tag{1}$$

Here, x denotes normalized importance, y denotes enhanced importance, and a controls the degree of enhancement. The horizontal shift places the inflection point at 0.5 because x lies between 0 and 1.

Next, we apply KDE with a Gaussian kernel to the enhanced importance distribution. Estimating the spatial density of locally clustered high-importance pixels reduces the effect of isolated low-importance pixels and represents neighboring high-importance pixels as continuous regions.

4) Temporal Integration: To cover the temporal domain, we generate short video sequences such that every frame of the interaction video sequence belongs to at least one short video sequence. We shift the sampling start frame by one frame at a time and continue sampling until the final sampled frame reaches the final frame of the interaction video sequence. Reordering the resulting importance maps chronologically makes relative changes in importance across consecutive frames easier to follow.

The second half of Fig. 1 illustrates this procedure. We place the importance maps from every short video sequence at their original frame positions. When several short video sequences contain the same source frame, we average the corresponding maps. This process integrates all maps into one chronologically ordered video sequence.

IV. Experiments

A. Pedestrian Group Video Sequences

We placed a camera 21.4 m above the ground to capture an overhead view of an outdoor parking lot. The camera recorded video at a resolution of 3840×2160 pixels and a frame rate of 30 fps. Twenty volunteers participated: 19 men and one woman, with a mean age of 22.6 ± 1.3 years.

We set the group size to two, the minimum number of people required for conversation. From the 20 participants, we randomly formed 52 unique pairs and recorded one pair at a time. We recorded every pair walking under each recording condition in Section II-B: active conversation, inactive conversation, and no conversation. Either member could initiate the conversation in the active and inactive conditions.

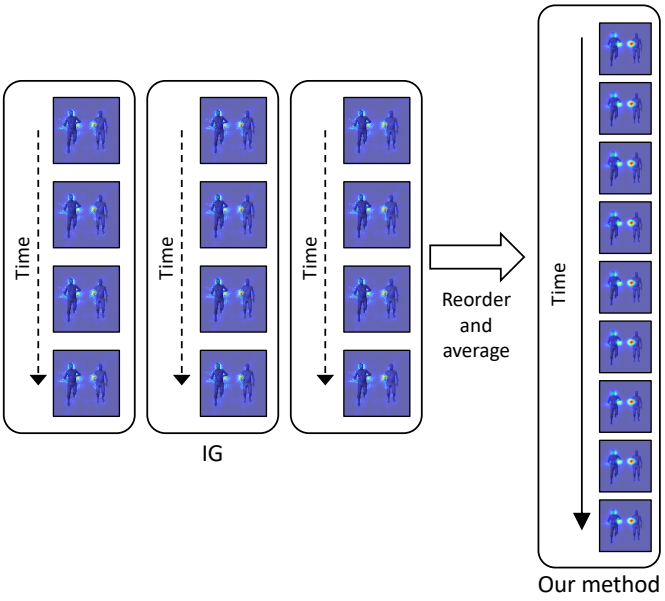

Fig. 3. Temporal integration of importance maps.

B. Visualization Parameters

We first generated interaction video sequences from the recorded pedestrian group video sequences and then generated short video sequences. We set the parameters in Section III-A to $h = 100$, $w = 100$, $I = 18$, and $F = 16$. Each short video sequence therefore had an array size of 100 pixels $\times$ 100 pixels $\times$ 16 frames $\times$ 3 RGB channels.

We visualized the spatiotemporal importance of pixels in each interaction video sequence using the method described in Section III-B. IG was computed for the model output corresponding to the class predicted by the C3D model, using an all-zero input sequence as the baseline and 50 steps with Gauss–Legendre quadrature. We set the sigmoid enhancement parameter to $a = 8$ and applied KDE with a Gaussian kernel to each enhanced importance map.

C. Effect of Spatial Enhancement

Figure 2 compares importance maps before and after spatial enhancement. Before spatial enhancement, low-importance pixels, shown in cyan to green, are finely scattered along pedestrian contours and across several body parts. This scattering obscures the body parts around which high-importance pixels, shown in yellow to red, are concentrated.

In the proposed method, the sigmoid function increases the contrast between high and low importance, and KDE groups locally concentrated high-importance pixels into continuous regions. These operations make regions with large attribution magnitudes around body parts such as the head and hands easier to identify visually.

D. Effect of Temporal Integration

Figure 3 compares importance maps before and after temporal integration. Before temporal integration, each short video

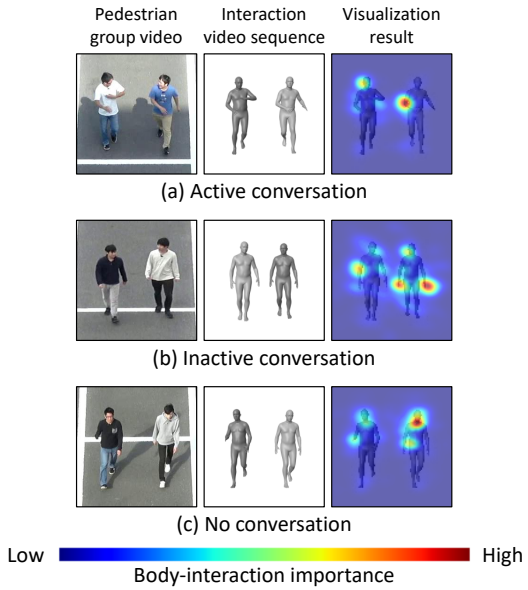

Fig. 4. Visualization results of body-interaction importance for interaction video sequences.

sequence presents its importance maps independently, so the temporal correspondence between sequences remains unclear. Temporal integration aligns the maps with their source frame indices and averages maps that correspond to the same frame.

This integration reduces discontinuities caused by boundaries between short video sequences and allows us to track changes in importance from the beginning to the end of an interaction video sequence. Averaging overlapping information also attenuates local variation that appears in only one short video sequence.

E. Discussion of Visualization Results

Figure 4 shows visualization results from the proposed method for each conversation activity label. We examine the visible body states and the body parts around which high-importance pixels are concentrated.

(a) Active conversation: The pedestrians raise their hands or elbows while gesturing and turn their heads and upper bodies toward one another. High-importance pixels are concentrated around the raised hands and elbows and around the mutually oriented heads and upper bodies. This pattern suggests that the C3D model may be sensitive to both gestures and mutual body orientation when predicting active conversation.

(b) Inactive conversation: The pedestrians keep their hands lowered and make no gestures, but they turn their heads toward one another. High-importance pixels are concentrated around the lowered hands and mutually oriented heads. The large attribution magnitudes around the heads suggest that the model may be sensitive to head orientation even when gestures remain limited.

(c) No conversation: The pedestrians make no gestures, keep their hands lowered, and orient their heads in the walking direction. High-importance pixels are concentrated around the lowered hands and the heads facing the walking direction. Inactive conversation and no conversation share the absence of gestures, while the attribution patterns around the heads differ depending on whether they face one another or the walking direction.

Overall, the visualized attribution patterns are consistent with our hypothesis that gestures and mutual body orientation may provide cues for conversation activity recognition. Future work will quantify these contributions through body-part masking and perturbation experiments.

V. Conclusion

This paper proposed an IG-based method for visualizing pixel-level spatiotemporal importance in interaction video sequences for conversation activity recognition in outdoor walking groups. Spatially, a sigmoid function and KDE suppress scattered low-importance pixels and clarify body parts around which high-importance pixels are concentrated. Temporally, chronological reordering and averaging of overlapping maps reveal continuous changes in importance across the interaction video sequence.

The visualization results showed large attribution magnitudes around raised hands and elbows and mutually oriented heads and upper bodies during active conversation, lowered hands and mutually oriented heads during inactive conversation, and lowered hands and heads facing the walking direction when no conversation occurred. These observations provide insights into physical states associated with the recognition decisions of the C3D model. These patterns may guide model refinement by motivating architectures or feature representations that explicitly encode hand and elbow motion as well as mutual head orientation. For example, future recognition models could incorporate pose-based features or attention mechanisms that emphasize hand, elbow, and head-orientation cues suggested by the visualization results. Future work will quantitatively evaluate visualization performance and the contribution of each body part. We would like to thank Dr. Michiko Inoue for their valuable suggestions during this research.

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] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3319–3328.

[3] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, 1992.

[4] F. Zanlungo, D. Brscic, and T. Kanda, "Pedestrian group behaviour analysis under different density conditions," *Transportation Research Procedia*, vol. 2, pp. 149–158, 2014.

[5] I. Chamveha, Y. Sugano, Y. Sato, and A. Sugimoto, "Social group discovery from surveillance videos: A data-driven approach with attention-based cues," in *Proceedings of the British Machine Vision Conference*, 2013, pp. 121.1–121.12.

[6] 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.