Detection of distant eye-contact using spatio-temporal pedestrian skeletons

Ryusei Hata[1], Daisuke Deguchi[1], Takatsugu Hirayama[1], Yasutomo Kawanishi[1,2], and Hiroshi Murase[1]

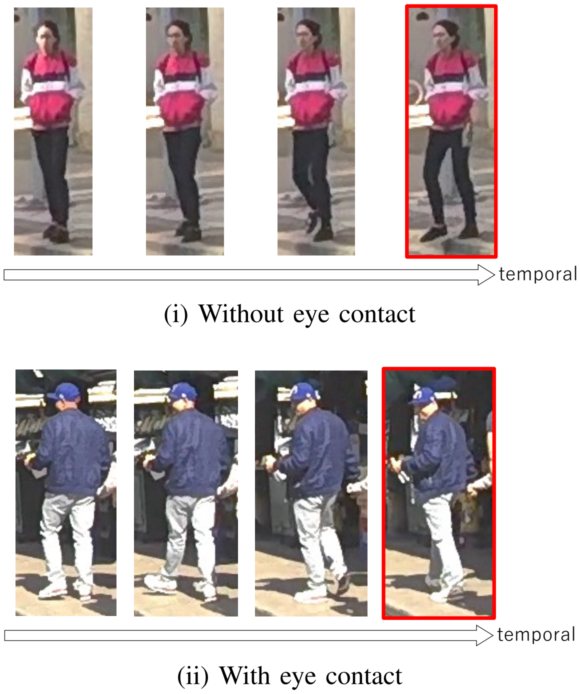
***Abstract*— Eye contact is an important factor in determining whether a pedestrian is aware of a vehicle. Most conventional eye contact detection methods rely on direct eye-gaze estimation based on eye measurements. This approach is difficult with distant pedestrians, especially in road environments. In contrast, the presence or absence of eye contact from a pedestrian can be determined based on information on their whole body and movement, such as their facial orientation, posture, and turn. In this study, we propose a method to detect eye contact from even distant pedestrians whose direction of gaze would be difficult to measure. The method captures the relationship between face and body information and its changes over time using spatio-temporal skeleton graph convolution. In an experiment using in-vehicle camera images, an accuracy of 88.6% was achieved. The results demonstrate the effectiveness of the proposed approach.**

I. INTRODUCTION

To prevent hazards and collisions while driving a vehicle, determining whether pedestrians are aware of its presence is important. Whether pedestrians direct their gaze at the vehicle, that is, whether they establish eye contact, is among the most important factors in determining whether they recognize the presence of a vehicle. Hence, to enable autonomous driving applications, the development of a technology that can accurately recognize the presence or absence of eye contact from pedestrians is expected. Such methods would provide a foundation for applications such as more detailed prediction of pedestrian behavior and situation-aware route planning.

Thus far, a great deal of research has been conducted on gaze estimation, a task closely related to eye contact detection. Most of the previous works on gaze estimation used face images taken close to a person's eyes as input and directly estimated their gaze based on eye measurements and appearance around the eyes [1]–[6]. Therefore, cameras must be located very close to the target person, and gaze estimation using these methods is not possible for pedestrians with distances from a road vehicle that would be consistent with normal traffic situations. Therefore, a method that does not rely on eye-gaze estimation is necessary to detect eye contact from distant pedestrians because resolution degrades according to distance.

In contrast, considering common driving situations, it may be noted that most drivers naturally consider temporal changes in face orientation and posture to determine whether a pedestrian is aware of the moving vehicle. Based on these findings, we aim to detect eye contact even for distant pedestrians for whom direct eye contact estimation is difficult.



(i) Without eye contact

(ii) With eye contact

Fig. 1. Example images of pedestrians with a similar face orientation

For example, let us consider the presence or absence of eye contact from a pedestrian at the time indicated by the red frame in Fig. 1. Face orientation is an important factor in identifying eye contact from a distant pedestrian. However, as illustrated in Fig. 1(i) and (ii), the presence or absence of eye contact may differ, even when the face orientations are similar. Considering body posture, we may note that the face and body of the person in Fig. 1(i) are oriented in the same direction, whereas the person in (ii) has adopted a posture that gradually turns towards the camera along with his height, i.e., he is in a ”twisted” state. Hence, we can infer that the person in (i) is not looking at the camera, and that the person in (ii) is. Furthermore, considering not only the frame shown in the red frame, but also the frames before it, we can see the movement of the person as they turn around. These movements would clarify whether the person is looking at the camera at the time of the red frame. That is, the presence or absence of eye contact can be estimated by considering face direction, body posture, and body movements in a complex manner.

In this study, face orientation, body posture, and their

[1] Nagoya University Graduate School of Informatics, Japan hatar@vislab.is.i.nagoya-u.ac.jp, {ddeguchi, takatsugu.hirayama, murase}@nagoya-u.jp

[2] Guardian Robot Project, R-IH, RIKEN, Japan yasutomo.kawanishi@riken.jp

movements are considered in terms of a sequence of skeletons. This feature is robust to differences in human appearance and has been widely used in research on action recognition. In particular, graph convolution, which connects keypoints of the sequence of skeletons as a graph structure, has attracted attention in recent years [8], [9]. This method is capable of recognizing complex motions because it considers not only the features of individual joint coordinates but also the spatial relations between connected joints and their temporal relations over frames.

Based on the above, we propose an eye contact detection method that captures characteristic temporal changes in the face and body using a sequence of skeletons represented by a graph structure. The two contributions of this paper are summarized as follows.

1) By representing face orientation, body posture, and their movements as a sequence of skeletons, the proposed method detects the eye contact of distant pedestrians whose direction of gaze cannot be measured directly with high accuracy.
2) We have clarified the effectiveness of graph convolution and the effect of keypoints (human joints) on eye contact detection using a sequence of skeletons.

II. RELATED WORK

A. Gaze estimation and eye contact detection using appearance features around the eyes

The task of gaze estimation is related that of eye contact detection, and has been studied in various ways. As an example, researches on gaze estimation by directly measuring the eyeball have been conducted [1]–[3]. These methods use pupil centre corneal reflection (PCCR), which combines an infrared light illuminator and an infrared camera, and are capable of highly accurate gaze estimation. However, there are some problems in practical applications, such as the need for special equipment as noted above. They must also be pre-calibrated, because infrared reflection is affected by individual differences.

In contrast, Baltrusaitis et al. proposed OpenFace, which detects facial landmarks in visible-light images to estimate gaze [4]. Although this method can estimate gaze from images only, it requires accurate detection of landmarks around both eyes and is limited by the orientation of the face and by distance.

In addition, Smith et al. proposed a method to estimate the presence of eye contact based on the appearance of the eyes [5]. By reducing the resolution of images taken from a distance of 2 m, this method achieved high estimation results, even at a resolution equivalent to 18 m. However, this approach has not been validated using actual images of people captured at a distance of 18 m. In addition, the dataset was collected under the condition that the head position was fixed on the apparatus, the face orientation was -30° in the yaw direction, and the pitch direction was constant, so it remains unclear whether the system would be able to handle sideways or backward-facing faces. In addition, the background of the images was plain, which presents another limitation.

Zhang et al. proposed a method for direct eye contact detection from human face images using CNNs [6]. However, in the training phase, they used facial landmarks obtained from OpenFace, and the training images were limited to those with clear facial landmarks. Therefore, the performance of the system may be expected to degrade for distant pedestrians with low resolution. In addition, face detection is required during the estimation phase.

As described above, most conventional methods have the following limitations: they require high-resolution images captured at close range, they have limitations on the image capture environment, and they require calibration for each subject and situation.

In contrast, in the present work, we consider a situation in which the face direction and posture of a pedestrian in the distance change over time, and some pedestrians may face away from the camera. The conventional methods are difficult to apply to these situations. In this study, we propose a novel method to detect eye contact with high accuracy even for distant pedestrians whose direction of gaze cannot be clearly observed in an unconstrained road environment.

B. Eye contact detection using human joints

Belkada et al. proposed a method to detect eye contact using a skeleton [7]. This method achieved highly accurate eye contact detection for distant pedestrians in an unconstrained environment. However, the method takes the coordinates of skeleton joints as input to a fully connected neural network, and does not explicitly consider joint relationships. In addition, because the method uses only a single frame, the motion information of the pedestrians is not considered. In contrast, the proposed method detects eye contact using a sequence of skeletons that considers the spatial and temporal connectivity of joints. The proposed method achieves highly accurate eye contact detection by considering the motion of the pedestrian in addition to the connection relations between their joints.

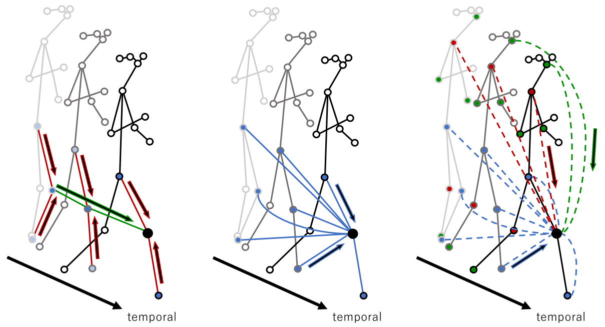
C. Action recognition using spatio-temporal skeleton graph convolution

In recent years, various action recognition methods based on spatio-temporal skeleton graph convolution have been proposed. ST-GCN [8] and its extended version of MS-G3D [9] are representative methods.

ST-GCN captures spatial features by performing graph convolution between keypoints on each frame, and captures temporal features by performing 1D-Conv of each keypoint in the temporal direction (Fig. 2 (a)).

In contrast, Liu et al. proposed G3D (Fig. 2(b)) that is a spatio-temporal skeleton graph convolution method connecting joints not only within but also between frames. While ST-GCN considers spatial and temporal elements separately, G3D can more directly capture the relationship between a joint and other joints at different times.

.



(a) ST-GCN (b) G3D (c) MS-G3D

Fig. 2. Example methods for graph convolution of a sequence of skeletons

Furthermore, Liu et al. extended G3D and proposed MS-G3D, which creates spatial graphs for each hop number from the keypoint of interest to perform multi-scale convolution. Thus, keypoints that are structurally distant from each other but have important meanings in recognition are connected (Fig. 2(c)). With its multi-scale receptive fields across both temporal and spatial dimensions, MS-G3D achieves high accuracy in the field of action recognition.

These methods realize action recognition that considers not only the spatial structure of the skeleton but also its temporal changes, i.e., motion.

III. METHOD

A. Outline

In this paper, we propose a method to detect the eye contact of a pedestrian by using a spatio-temporal skeleton graph convolution with keypoints (human joints) of each frame.

Fig. 3 is a sequence of skeletons of the pedestrian image series in Fig. 1. The blue color represents the keypoints, the solid red line represents the spatial connections between the keypoints, and the dashed green line represents the temporal correspondence between the same keypoints (the connection of the left ear and left shoulder is shown here as an example). Comparing the keypoints at the timing of the red boxes in Fig. 3(i) and (ii), it may be observed that the positional relationships of the keypoints from the neck upwards was similar. In contrast, the keypoints of the body below the neck differ significantly, and it may be observed that the face and body of the person in (i) are turned in the same direction, whereas the posture of the person in (ii) gradually turns toward the camera from the feet to face, i.e., the body posture is in a ”twisted” state. Thus, we can see that the face orientation and body posture of the pedestrian are sufficiently represented by keypoints. In addition, if we focus on the frames before the timing of the red frame, we can also observe the movements of a pedestrian’s body, such as turning and twisting, by changing the coordinate of each keypoint. Based on these features of a sequence of skeletons of the pedestrian, the proposed method detects eye contact, considering the facial orientation, body posture, and motion of the pedestrian.

The features of each keypoint are obtained using the posture estimator described below and are represented by 3D feature vectors with $x-y$ coordinates and estimated confidence c, respectively. However, simply using the feature values of these keypoints does not allow us to explicitly consider the spatial connectivity between them (solid red line in Fig. 3) and temporal correspondence between the same keypoints (dashed green line in Fig. 3). Therefore, in the proposed method, these connection relations are represented using a graph structure inspired by the researches on action recognition [8], [9]. Specifically, we represent spatio-temporal skeleton graph using an adjacency matrix that represents the spatial and temporal connectivity of keypoints extracted from each frame. A graph convolution is performed on a sequence of skeletons represented as a graph to account for the spatio-temporal relationship of keypoints. This enables us to detect eye contact from a pedestrian more strongly by capturing their face direction, body posture, and motion compared with the simple use of 3D keypoints.

B. Representation of spatio-temporal skeleton graph convolution

The set of keypoints extracted from T frames is represented as

$$\chi = \{\mathbf{x}_{t,n} \in \mathbb{R}^3 \mid t, n \in \mathbb{Z}, 1 \leq t \leq T, 1 \leq n \leq N\}. \quad (1)$$

The spatio-temporal skeleton graph is represented by using the adjacency matrix $\mathbf{A} \in \mathbf{A}^{(T\times N)\times(T\times N)}$, where N is the number of keypoints of the human skeleton in each frame. The set of keypoints χ and the spatial connectivity of the adjacency matrix $\mathbf{A}$ corresponds to the representation of the “skeleton” on a frame-by-frame basis, and the temporal connectivity of $\mathbf{A}$ results in the representation of “a sequence of skeletons”. In addition, χ can be represented by the tensor $\mathbf{X} \in \mathbb{R}^{T\times N\times 3}$, where $\mathbf{x}_{t,n} = \mathbf{X}_{t,n}$ is a 3-dimensional feature vector.

A spatio-temporal skeleton graph convolution is performed by a neural network f_p that uses the keypoint set $\mathbf{X}$ and adjacency matrix $\mathbf{A}$ to represent a graph of keypoints connected over T frames, as shown in Fig. 2. The spatio-temporal skeleton graph convolution is represented by the following equation using the weight parameter θ.

$$\mathbf{x}_p = f_p(\mathbf{X}, \mathbf{A}, \theta) \quad (2)$$

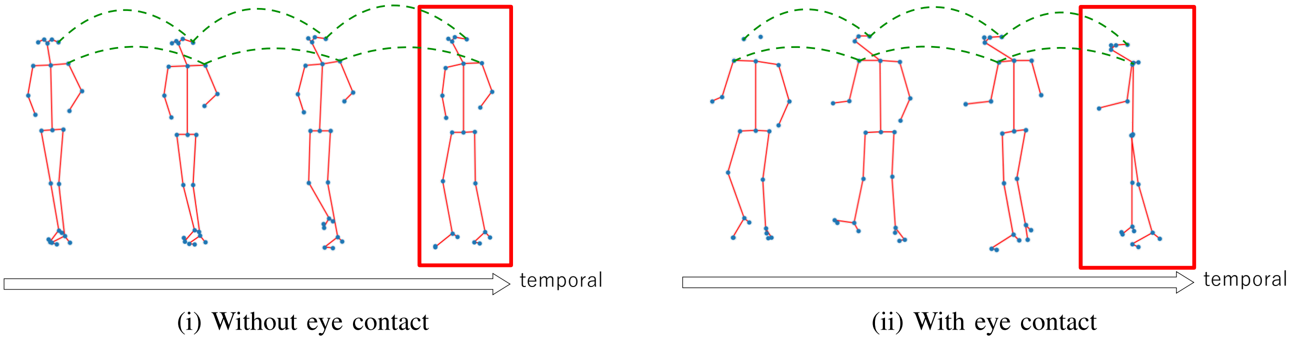
This converts the sequence of skeleton graphs into a feature $\mathbf{x}_p$, which considers the spatial connectivity between joints and the temporal relationship between frames. The obtained $\mathbf{x}_p$ is used to classify images into the presence and absence of eye contact for the T-th frame.

C. Processing steps

Fig. 4 shows the process steps of the proposed method.

1) Perform the pedestrian detection on the image sequence consisting of T frames captured by the in-vehicle camera (hereinafter referred to as the in-vehicle

.



(i) Without eye contact

(ii) With eye contact

Fig. 3. Skeleton series representation of Fig. 1

.

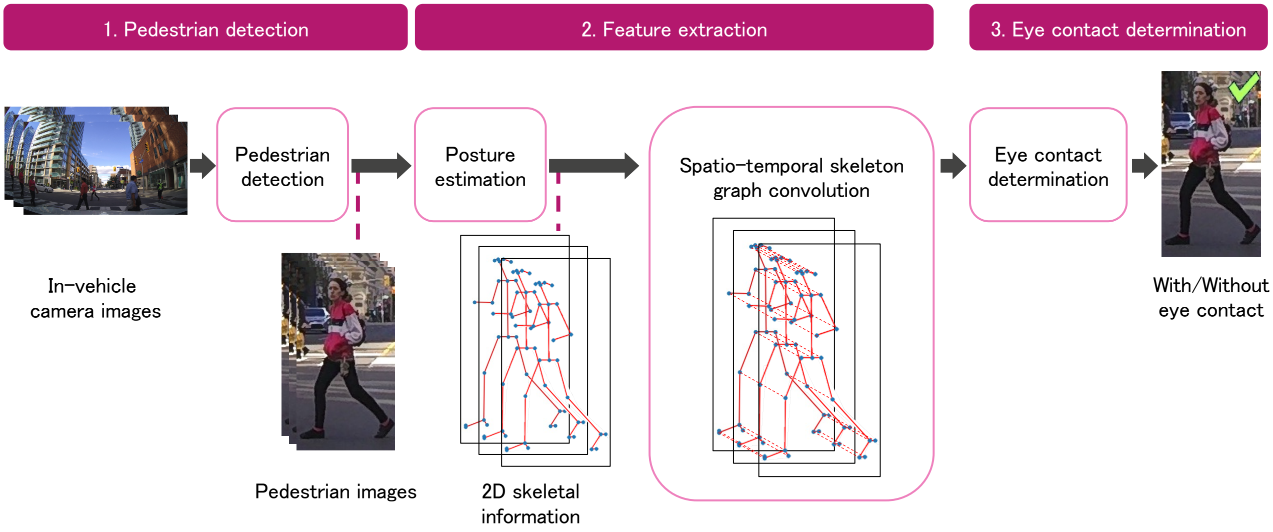


Fig. 4. Processing steps of the proposed method

camera image sequence) and obtain the pedestrian area image (hereinafter referred to as the pedestrian image).

2) Perform the skeletal estimation on each of the pedestrian images and obtain 2D skeletal information (2D coordinates of each keypoint and their confidence levels).

3) Perform the spatio-temporal skeleton graph convolution as explained in III-B to obtain the feature value $\mathbf{x}_p$.

4) Input the feature value $\mathbf{x}_p$ to the eye contact classifier to obtain the determined result of the presence or absence of eye contact for the T-th frame.

The details of each process and its implementation are described below.

D. Pedestrian detection

A pedestrian detector is used to detect the pedestrian area for each image in the in-vehicle camera image series.

E. Extraction of keypoint features

Using a skeletal estimator, we obtain keypoint features (C-dimensional feature vectors) for N keypoints from each pedestrian image. In this work, we used 25 keypoints, as shown in Fig. 5. The feature value of keypoint n at time t is then obtained as a 3D feature vector $(\widetilde{x}_{t,n}, \widetilde{y}_{t,n}, \widetilde{c}_{t,n})$, whose elements are the x- and y-coordinates in sub-pixel units with the origin at the upper-left corner of the pedestrian image and the confidence level c of its estimation. Here, $N = 25$ and $C = 3$ dimensions. $\widetilde{c}_{t,n}$ is a real number in the range $[0, 1]$. If $\widetilde{c}_{t,n} = 0$, it is assumed that no keypoint was detected and the values of $\widetilde{x}_{t,n}$ and $\widetilde{y}_{t,n}$ are both zero (missing keypoints). Then, the feature vectors of all the keypoints are combined, and $\widetilde{\mathbf{p}}_t = (\widetilde{x}_{t,0}, \widetilde{y}_{t,0}, \widetilde{c}_{t,0}, ...\widetilde{x}_{t,24}, \widetilde{y}_{t,24}, \widetilde{c}_{t,24})$ is obtained as a 75-dimensional feature vector.

In this case, because the scale of the coordinate information of $\widetilde{\mathbf{p}}_t$ differs for each skeletal graph, it is normalized so that the sitting height L of the person in the image is aligned among the persons. The sitting height is defined as

$$L = \sqrt{(\widetilde{x}_{t,12} - \widetilde{x}_{t,5})^2 + (\widetilde{y}_{t,12} - \widetilde{y}_{t,5})^2}. \tag{3}$$

and the keypoint n is normalized using the following equa-

.

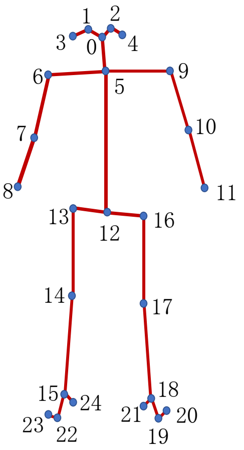


Fig. 5. Keypoints used in this study

tions.

$$\widehat{x}_{t,n} = \begin{cases} \frac{\widetilde{x}_{t,n}-\widetilde{x}_{t,12}}{L} & (\widetilde{c}_{t,n} > 0) \\ 0 & (\widetilde{c}_{t,n} = 0), \end{cases} \tag{4}$$

$$\widehat{y}_{t,n} = \begin{cases} \frac{\widetilde{y}_{t,n}-\widetilde{y}_{t,12}}{L} & (\widetilde{c}_{t,n} > 0) \\ 0 & (\widetilde{c}_{t,n} = 0), \end{cases} \tag{5}$$

$$\widehat{c}_{t,n} = \frac{\widetilde{c}_{t,n} - 0.5}{0.5} = 2\widetilde{c}_{t,n} - 1. \tag{6}$$

The resulting normalized features are denoted by $\widehat{\mathbf{p}}_t$, and the set of features extracted from T frames is denoted by $\widehat{\mathbf{p}} = \{\widehat{\mathbf{p}}_t | t \in \mathbf{T}\}$.

F. Spatio-temporal skeleton graph convolution

MS-G3D (Fig. 2(c)) was used as f_p of the neural network for spatio-temporal skeleton graph convolution. Specifically, $\widehat{\mathbf{p}}$ obtained in III-E is input to f_p, and then the feature $\mathbf{x}_p = f_p(\widehat{\mathbf{p}}, \theta)$ is calculated.

G. Eye contact determination

The feature value $\mathbf{x}_p$ obtained in Section III-F is used in the eye contact classifier, which identifies the two classes with and without eye contact in the last frame of the input of a sequence of skeletons and produces the final output. The eye contact classifier consists of a single-layer, fully connected neural network.

IV. Experiments

A. Dataset

In this experiment, the Pedestrian Intention Estimation dataset (hereafter referred to as the PIE dataset) [12] was used as a source of in-vehicle camera images. The PIE dataset contains 1,842 pedestrians in total, and it includes information such as pedestrian ID, whether the pedestrian is looking at the camera, bounding rectangles, and degrees of obstruction. In this experiment, the annotation of whether the pedestrian is looking at the camera was used as the ground truth for the presence or absence of eye contact. We used the rectangular information in the PIE dataset as the pedestrian detection result instead of the output of a pedestrian detection process. BODY_25 model of OpenPose [11] was used as the skeletal estimator. To simplify the problem, only pedestrian images that satisfied the following conditions were used for the evaluation.

1) Less than 25% of pedestrians were occluded
2) The vertical size of the cropped pedestrian image was greater than 150 pixels
3) The rectangle of the cropped pedestrian image did not exceed the frame of the original in-vehicle camera image $(1,920 \times 1,080)$

In addition, we extracted pedestrian images for the experiment to ensure that the number of data with and without eye contact would be equal. Details of the dataset are shown in Table I.

TABLE I
BREAKDOWN OF EXPERIMENTAL DATA

Type of data	Number of data
Train	21,924
Validation	2,126
Test	12,546

TABLE II
COMPARISON WITH THE CASE WITHOUT THE GRAPH (AVERAGE ACCURACY OF 5 TRIALS)

Method	Usage model	1 frame	10 frames
Comparison method [7]	FCN	85.5%	87.8%
Proposed method	MS-G3D	86.5%	**88.6%**

B. Evaluating the effectiveness of using spatio-temporal skeleton graph convolution

We conducted an experiment to verify the usefulness of spatio-temporal skeleton graph convolution in an eye contact detection task.

In this experiment, we compared the two methods presented in Table II. In this experiment, we used the face and torso of the body (6, 9, 12, 13, and 16 in Fig.5) as input keypoints. The method used for comparison (1 frame) is the same as that of Belkada et al. [7], which does not use a graph representation of the skeleton and inputs all keypoints into a fully connected neural network (FCN). The FCN consisted of three layers. For each method, the cross-entropy loss and AdamW [10] optimizer were used to train the model.

For each method, we performed eye contact detection using only the skeletal information of a single frame and using a sequence of skeletons of 10 frames. A single frame FCN corresponds to the method proposed by Belkada et al. [7]. In this experiment, we performed five trials with different seeds, and compared the average accuracy.

The experimental results are shown in Table II. Compared with the comparison method, the proposed method was more accurate for both single and multiple frames of input. The

TABLE III

COMPARISON OF ACCURACY BY FEATURES CONSIDERED (AVERAGE ACCURACY OF 5 TRIALS)

Method	Feature to use	1 frame	10 frames
Method 1	face only	85.6%	87.6%
Method 2	face + body	86.3%	87.6%
Method 3	face + body (torso only)	86.5%	**88.6%**
Method 4	face + body (torso + knees only)	86.4%	88.2%

proposed method achieved an 88.6% accuracy rate, compared to 85.5% for the existing method(a single frame FCN).

C. Ablation study: examining the effectiveness of using the face, body, and movement

The following experiment was conducted to evaluate the effect of capturing time-series changes in facial and body features.

In this experiment, we compared the four methods shown in Table III. Method 1 was an eye contact detection method using six keypoints on the face (0, 1, 2, 3, 4, and 5 in Fig. 5). In contrast, Method 2 used 25 keypoints on the face and body. Method 3 used only 11 keypoints of the torso (6, 9, 12, 13, and 16 in Fig. 5) and face. Method 4 used only 13 keypoints of the torso and knees (6, 9, 12, 13, 14, 16, and 17 in Fig. 5) and face. For each method, we performed eye contact detection using only the skeletal information of a single frame and using a sequence of skeletons of 10 frames. In this experiment, we performed five trials with different seeds, and compared the average accuracy.

The experimental results are shown in Table III. For all the methods, we confirmed that the correctness rate was higher when 10 frames were used than when only a single frame was used.

Compared with Method 1, which uses only the keypoints of the face, Methods 2, 3, and 4, which add the body features, showed an overall improvement in accuracy. In contrast, compared to Method 2, which used all the keypoints of the body, Method 3, which uses only the torso, and Method 4, which uses only the torso and knees, improved the accuracy.

Method 3 with 10 frames of input had the highest accuracy and also improved the accuracy by 3.0% over Method 1 with a single frame of input.

V. DISCUSSION

A. Useful body features

As shown in the table of experimental results in Section IV-C, Method 3, which included the torso of the body, exhibited the highest accuracy. The accuracy of Method 3 was also higher than that of Method 2, which used the whole body, and Method 4, which used the torso and both knees as body features. One reason for this is that the positions of the keypoints of the arms and legs vary significantly depending on pedestrians' gait conditions, which may have caused noise when judging eye contact detection. Fig. 6 shows an example of pedestrians whose eye contact was correctly determined by Method 3, while they were incorrectly determined by Methods 2 and 4. It may be observed that the positions of the keypoints of these pedestrians were similar when focusing on the face and torso. However, when we focus on the arms and legs, the positions of the keypoints differ greatly among walkers. In the case of the arms, the keypoint of the wrist is higher than that of the elbow in (i) and (iii), whereas it is lower in (ii). As for the legs, the right leg is forward in (i), while the left leg is forward in (ii). In addition, the person in (iii) had both legs together, which does not necessarily correlate with eye contact. Therefore, we believe that only the torso part is sufficient to consider the torsion of the body, and data on the arms and legs are mainly regarded as noise.

B. Effectiveness of using graph

As shown in the table of experimental results in Section IV-B, the proposed method with the graph is more accurate than the comparative method without.

Fig. 7 shows an example of a pedestrian image (upper body only) where eye contact was incorrectly determined by the FCN method and correctly determined by the MS-G3D method for a 10-frames sequence of skeletons as input. In all these examples, the positions of the keypoints of the face change because of turning around. The graph structure explicitly represented the amount of change in the position of each keypoint between frames because the convolution was performed by connecting the same keypoints in the time direction.

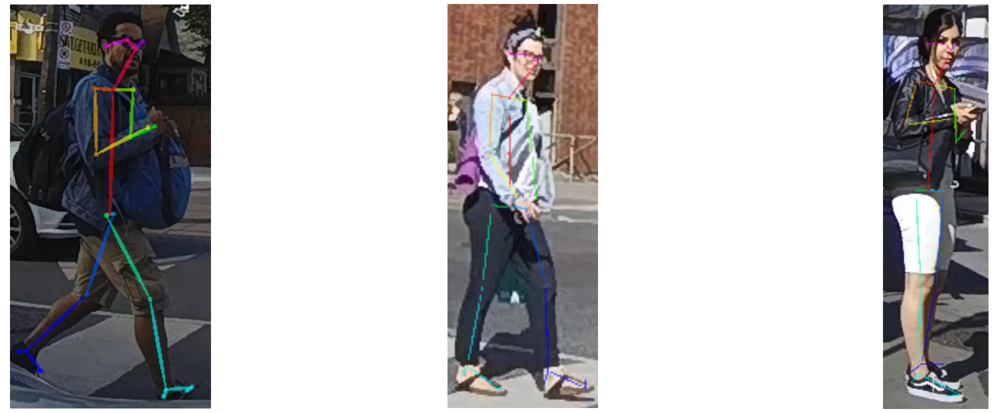
These results show that the proposed method, which uses a graph structure, was able to capture the changes in motion better than the existing method used for comparison.

C. Considerations for using a sequence of skeletons

As may be observed from the experimental results in Section IV-C, the accuracy was higher when multiple frames were input than when a single frame was input. In Method 3 (face + torso), Fig. 8 shows examples in which the presence or absence of eye contact was incorrectly determined for a single frame, but correctly determined for 10 frames (upper body only). In the example in Fig. 8(i), it may be observed that the presence of eye contact can be correctly determined by capturing motion. Also, Fig. 8(ii) shows an example in which skeletal estimation failed in some of the input frames. In this way, we believe that the spatio-temporal skeleton graph convolution of multiple frames improves robustness against errors in skeletal estimation, in addition to considering motion.

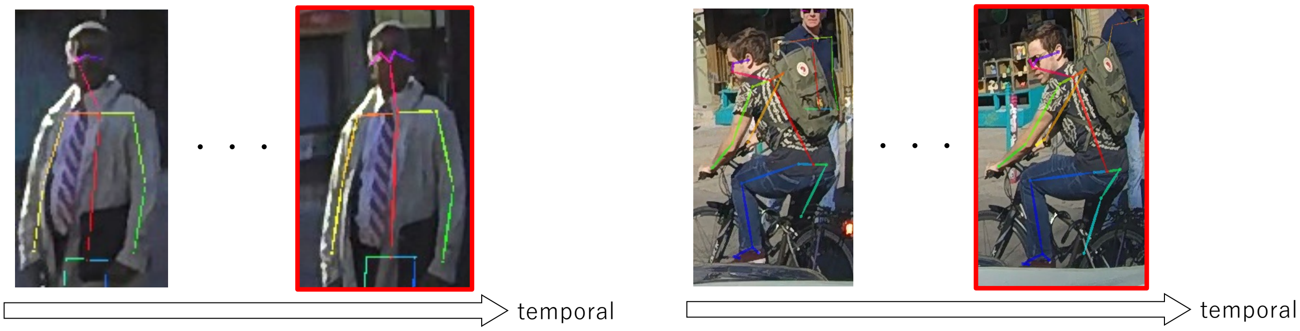
VI. CONCLUSIONS

In this study, we propose a method to detect eye contact from a pedestrian using spatio-temporal skeleton graph convolution, considering facial and body features and movements. In the proposed method, the pedestrian image is first extracted from the in-vehicle camera image using the pedestrian detector, and the keypoint features are obtained using the skeleton estimator. Then, a sequence of skeletons is convolved with an adjacency matrix representing the graph



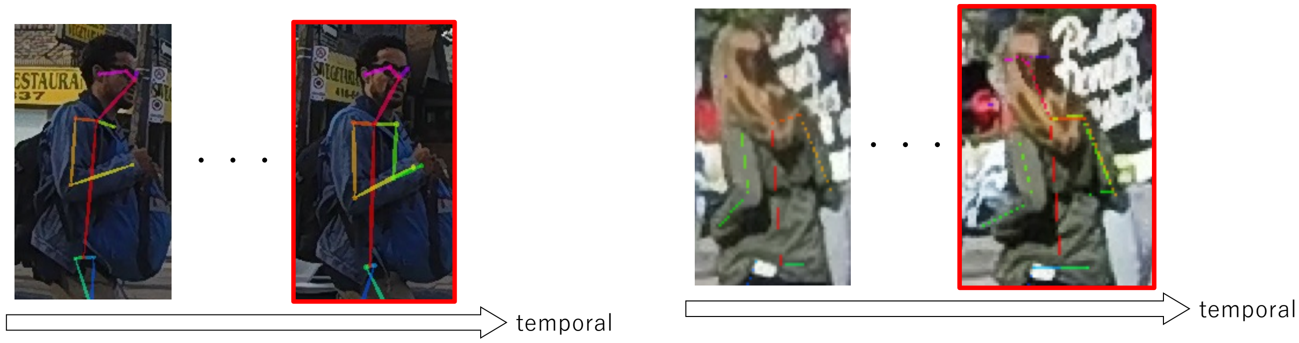
(i) With eye contact (ii) With eye contact (iii) With eye contact

Fig. 6. Examples of correct eye contact detection when body features are focused only on the torso.



(i) With eye contact (ii) With eye contact

Fig. 7. Examples of correct eye contact detection when using a graph



(i) An example of capturing sufficient motion. (ii) An example of correct eye contact detection even if the posture estimation fails partially

Fig. 8. Examples of correct eye contact detection when using a sequence of skeletons of multiple frames

structure and keypoint features to obtain spatio-temporal features. The features thus obtained are used to determine the presence or absence of eye contact.

We conducted an evaluation experiment using the PIE dataset to verify the effectiveness of considering face, body, and motion information, as well as the effectiveness of using spatio-temporal skeleton graph convolution. The results of these experiments indicated that eye contact detection is possible even for pedestrians who are distant and whose gaze is difficult to measure directly. The proposed method achieved an 88.6% accuracy, compared to 85.5% for the existing method (a single frame FCN).

Future work may include the investigation of keypoints that are particularly important in determining the presence or absence of eye contact, and the consideration of the context surrounding the pedestrian such as the presence and location of other vehicles and pedestrians and road environment such as crosswalks.

Acknowledgements This work was partially supported by JSPS Grant-in-Aid for Scientific Research 17H00745. The computation was carried out using the supercomputer ”Flow” at Information Technology Center, Nagoya University.

References

[1] C. Hennessey, B. Noureddin, and P. Lawrence, “ A single camera eye-gaze tracking system with free head motion, ” Proceedings of the 2006 symposium on Eye tracking research & applications, pp.87–94, Jan. 2006.

[2] D.H. Yoo and M.J. Chung, “A novel non-intrusive eye gaze estimation using cross-ratio under large head motion, ” Computer Vision and Image Understanding, vol.98, no.1, pp.25–51, Apr. 2005.

[3] K. Egawa, R. Yamamoto, and T. Nagamatsu, “ Development of Eye-tracking Volume Simulator for Corneal Reflection Method and Its Applications for Multi-user Gaze Interaction Systems, ” IPSJ Journal, vol.55, no.11, pp.2476–2486, Nov. 2014.

[4] T. Baltrusaitis, A. Zadeh, Y.C. Lim, and L. Morency, “Openface 2.0: Facial behavior analysis toolkit,” Proc. of the 13th IEEE International Conference on Automatic Face Gesture Recognition, pp.59–66, May 2018.

[5] B.A. Smith, Q. Yin, S.K. Feiner, and S.K. Nayar, “Gaze locking: Passive eye contact detection for human-object interaction,” Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, pp.271–280, 2013.

[6] X. Zhang, Y. Sugano, and A. Bulling, “Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery,” Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, pp.193–203, 2017.

[7] Y. Belkada, L. Bertoni, R. Caristan, T. Mordan, and A. Alahi, “Do Pedestrians Pay Attention? Eye Contact Detection in the Wild,” arXiv preprint arXiv:2112.04212, 2021.

[8] S. Yan, Y. Xiong, and D. Lin, “ Spatial temporal graph convolutional networks for skeleton based action recognition, ” Proc. of the 32nd AAAI Conference on Artificial Intelligence, pp.7444-7452, Feb. 2018.

[9] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, “ Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition, ” Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.143-152, June. 2020.

[10] I. Loshchilov and F. Hutter, “Fixing weight decay regularization in adam,” Proc. of the Ninth International Conference on Learning Representations, pp.1–14, Feb. 2018.

[11] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.43, no.1, pp.172–186, Jan. 2021.

[12] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, “PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction,” Proc. of the 2019 IEEE/CVF International Conference on Computer Vision, pp.6261–6270, Oct. 2019.