Detection of Texting while Walking in Occluded Scenarios Using Variational Autoencoder

Hayato Terao[1], Jiaxu Wu[2], Hajime Asama[2], Qi An[1] and Atsushi Yamashita[1]

***Abstract*— Texting while walking is a common behavior exhibited by pedestrians. While several studies explored the detection of texting while walking, the influence of occlusions were neglected. In this paper, we propose an image-based method which utilizes a pre-trained Variational Autoencoder. The proposed method takes sequence of 2D coordinates of upper body key points of the pedestrians as input, encodes the data into a 2D latent space, and uses the encoded data to distinguish text walkers from normal pedestrians. The proposed architecture enables the model to extract meaningful features from occluded data. Results of ablation test and comparison with a previous method revealed that the proposed architecture is successful in identifying text walkers even under heavy occlusion, outperforming a previously proposed method.**

I. INTRODUCTION

Mobile robots have become ubiquitous across various domains, and their presence in our daily lives is steadily growing. From restaurants and shopping centers to streets and beyond, mobile robots are becoming a common sight in public spaces. With the increasing likelihood of human-robot interaction, it is imperative for mobile robots to comply with safety requirements, such as avoiding collisions between mobile robots and pedestrians.

One concerning behavior observed among pedestrians is texting while walking, herein referred to as "text walkers", as depicted in Fig. 1. Over 7% of pedestrians are observed to be engaged in texting while walking [1], coinciding with the escalation in pedestrian injuries associated with the use of mobile phones [2]. Text walkers exhibit distinct behaviors such as fixating on their phone screens and paying less attention to their surroundings, resulting in an elevated risk of collisions with other pedestrians or obstacles [3]. Moreover, text walkers inadvertently cause slowdowns in pedestrian flow and compel nearby pedestrians to make sudden turns to avoid collisions [4], introducing an additional risk of collision due to unexpected behaviors. Thus, it is crucial for a mobile robot to identify text walkers ahead of time to minimize the risk of collisions with pedestrians.

Existing studies delved into the identification of text walkers using sensors on mobile robots [5]–[7]. However, none of these studies explicitly tackled the challenge of detecting text walkers in occluded scenarios. Occlusion, wherein a pedestrian's body is partially or entirely concealed from view, manifests over 70% of pedestrian situations [8] and its effect cannot be undermined. Some unique features of text walkers may be undetectable under occlusion and previously proposed methods may not function in real-world situations. Consequently, it is imperative to develop an approach capable of detecting text walkers even amidst occluded scenarios.

Fig. 1. Example of Normal Pedestrian (left) and Text Walker (right).

This paper introduces a novel image-based machine learning approach for identifying text walkers. The proposed method utilizes a pre-trained Variational Autoencoder (VAE) [9] to supervise the feature extraction, enabling a meaningful feature extraction from input data even with substantial occlusions. The resulting features are then used to classify the pedestrian into either normal pedestrian or text walker.

The subsequent sections of the paper are organized as follows: Section II describes the related works, followed by Section III which illustrate the proposed method, detailing its mechanism. Section IV presents the data collection and experiment of the proposed method, and the results are presented and analyzed in Section V. Finally, the conclusion and future works are given in Section VI.

II. RELATED WORKS

A. Detection of Text Walkers

Several existing works have addressed the detection of text walkers by using images [5], [6] or point cloud data of pedestrians [7]. For instance, Kumamoto et al. [5] analyzed the body pose of pedestrians from an RGB image to categorize their actions into normal walking, texting while walking, or talking on a phone while walking. Their method achieved a classification accuracy of 89.6% for pedestrians in real-world environments. However, images of occluded pedestrians were ignored in their research.

[1]H. Terao, Q. An and A. Yamashita are with the Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8563, Japan {terao, anqi, yamashita}@robot.t.u-tokyo.ac.jp

[2]J. Wu and H. Asama are with the Department of Precision Engineering, Graduate School of Engineering, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8563, Japan {wujiaxu, asama}@robot.t.u-tokyo.ac.jp

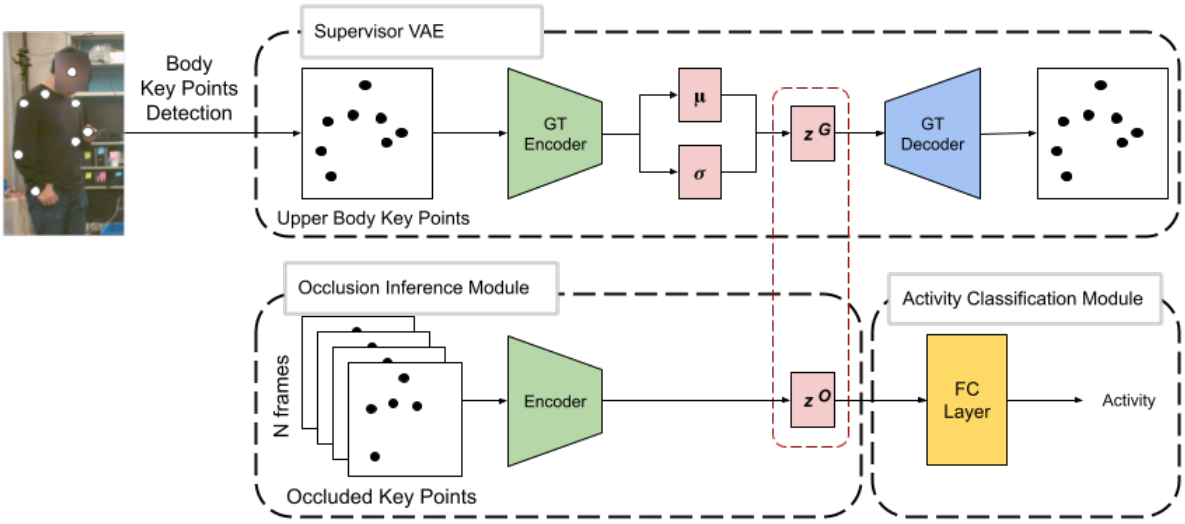

Fig. 2. Overview of the Proposed Method.

In contrast to using an image, Wu et al. [7] proposed an approach of using a LiDAR sensor and analyzed the point cloud pattern of a pedestrian to distinguish between normal pedestrians and text walkers. A maximum F1 score of 0.44 is reported for their approach, suggesting challenges in classifying text walkers compared to image-based methods.

One common aspect among these methods is the reliance on a single frame of data to discern the pedestrian's activities. Although their method may function in ideal situations where all necessary pedestrian information is captured in a single frame of data, the limitation becomes pronounced when the data is occluded causing data insufficiency, as the extraction of features might be unsuccessful.

Another study leveraged data obtained from pedestrian's smartphone and a body-worn accelerometer to detect text walkers [10]. While this approach offers potential robustness against occlusion, its real-world viability is questionable, as it necessitates pedestrians to provide their smartphone data and wear additional sensor. Hence, the detection of text walkers in occluded scenarios remains unresolved in practical terms.

B. Perception in Occluded Scenarios

A study by Dollar et al. [8] revealed the prevalence of occlusions among pedestrians on streets. A video taken with an RGB camera mounted on a vehicle was analyzed, and it was discovered that 71% of pedestrians are occluded in at least one frame, and 79% of them are heavily occluded, indicating over 40% of their body were hidden. Occlusions pose a challenge, as the performance of pedestrian detection substantially decreases in the presence of occlusions [11].

One approach explored the use of image inpainting [12] to detect the pose of a bus driver partially occluded by a steering wheel [13]. Images of the occluded body parts of the bus driver such as the arms were reconstructed with image inpainting. The reconstructed image was used to detect the body key points of the bus driver, which demonstrated improvements in body key points detection. While this method appears to be effective for mutually occluded person where an occlusion is caused by other objects or people, it may exhibit limited performance in the case of self-occlusion. Self-occlusion is the case in which the occlusion is caused by the person themselves, such as the person's hand being hidden by their body. In such situation, image inpainting will attempt to reconstruct the background instead of the concealed body parts, limiting the application of this method.

Instead of focusing on one pedestrian, another approach paid attention to observing all pedestrians in the scene [14]. This approach is grounded in the concept that one can anticipate the presence of a pedestrian or an obstacle by observing the behaviors of other pedestrians. For instance, if a pedestrian suddenly decelerates, it may suggest the existence of a hidden obstacle in their path. Several studies have demonstrated the effectiveness of this approach in predicting the presence of fully occluded pedestrians or obstacles [14]–[16]. However, this approach is limited to detecting the presence of a potential pedestrian, and not specifically identifying their activities.

III. PROPOSED METHOD

A. Problem Setting

We consider a scenario in which a pedestrian's body is only partially visible to a mobile robot. Our approach involves capturing pedestrian images using an RGB camera positioned at a height of 1 meter from the ground, replicating a view from a mobile robot, and using the image to detect and classify pedestrians into normal pedestrian or text walker. The choice of an RGB camera is driven by the human's ability to identify text walkers even amidst partial occlusions, just through visual observation.

B. System Overview

To address the challenge of detecting text walkers under occluded scenarios, we propose a machine learning-based approach outlined in Fig. 2. In contrast to the previous

image-based methods [5], [6], our approach takes a sequence of pedestrian body key points as input. It is hypothesized that the sequential data-based approach imparts an increased robustness against temporary occlusions, as the missing information may be inferred from adjacent frames of data.

We define temporary occlusion as the case when all body key points of a pedestrian is observed at least once during the input sequence. Oppositely, persistent occlusion denotes the case where at least one key point remains unobservable throughout the entire input sequence.

To fortify our model against persistent occlusion, we employ an architecture introduced in [16]. This architecture incorporates a VAE pre-trained with unoccluded pedestrian data, using it as a supervisor to train the feature extraction process of occluded data which we call the occlusion inference module (OIM). The OIM takes a sequence of potentially occluded body key points and outputs a latent representation of the pedestrian pose. In other words, the OIM extracts pedestrian features while robustly accommodating occlusions.

Finally, the activity classification module (ACM) takes the features from the OIM to classify the pedestrian into normal pedestrian or text walker.

C. Supervisor Variational Autoencoder

The role of the supervisor VAE is to oversee the training phase of the feature extraction by the OIM. Its involvement is exclusive to the training phase and does not extend to testing, as it requires ground truth data of body key points which are unavailable during testing. Furthermore, only the encoder of the VAE is used for supervising the training phase, as the encoder is responsible for extracting the features of the input data into a latent space. The VAE is pre-trained with unoccluded data of pedestrian body key points, encoding it into a latent representation and decoding it back to the original data.

We follow the standard procedure of training a VAE, employing the ELBO loss [9] encompassing L2 reconstruction loss and Kullback-Leibler divergence (KL loss). The reconstruction loss monitors how close the coordinates of the reconstructed pedestrian body key points R_t is to that of the original key points P_t, and is calculated as:

$$L_{recon} = ||P_t - R_t||_2^2. \tag{1}$$

The KL loss represents how close the probabilistic distribution of the latent vector generated by the encoder is to that of Gaussian distribution. Its role is to control how distributed the latent vectors are, and is calculated by the following equation:

$$L_{kl} = -\beta(1 + \log\sigma^2 - \mu^2 - \sigma^2), \tag{2}$$

where μ and σ are the mean and standard deviation of the distribution of the latent representations, respectively. β is the KL coefficient [17], a hyperparameter for adjusting the balance between L_{recon} and L_{kl}.

Finally, the ELBO loss is computed as the sum of the two losses: L_{recon} and L_{kl}.

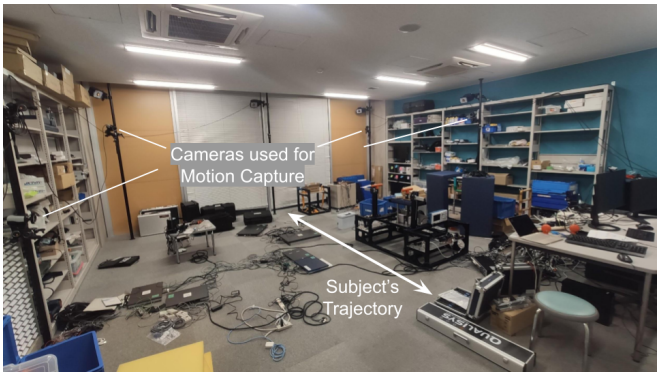

Fig. 3. Experimental Setup.

D. Feature Extraction with Occlusion Inference Module

The role of the OIM is to take a sequence of occluded pedestrian body key points $O_{t-T:t}$ with the aim of deriving robust features that remain resilient to occlusions. The feature extraction is overseen by the supervisor VAE by employing encoding loss $L_{encoding}$ computed as:

$$L_{encoding} = ||z_{vae} - z_{oim}||_2^2, \tag{3}$$

where z_{vae} is the latent representation of P_t outputted by the supervisor VAE, and z_{oim} is the latent representation of $O_{t-T:t}$ by the OIM.

This enables the OIM to extract features from observed key points, encompassing information of the occluded key points, as it would do for fully observable data.

E. Text Walker Detection with Activity Classification Module

The activity classification module processes the input z_{oim} and performs the task of classifying pedestrians into two categories: normal pedestrians or text walkers. This module is trained by utilizing categorical cross-entropy loss L_{cat} which quantifies the discrepancy between the predicted and actual labels assigned to the pedestrians.

Importantly, the OIM and the activity classification module undergo simultaneous training within a unified stage. This approach ensures that the model's optimization is intricately aligned with the accurate determination of pedestrian activities. The loss is calculated as shown below:

$$L = w_{encoding} \cdot L_{encoding} + (1 - w_{encoding}) \cdot L_{cat}, \tag{4}$$

where $w_{encoding}$ is a coefficient applied to take a balance between the encoding loss and cross-entropy loss.

IV. EXPERIMENT

A. Data Collection and Preprocessing

The experimental setup is represented in Fig. 3 Three individuals, each having experience of texting while walking, participated in the data collection. Their task involved traversing a room under two conditions: walking in a regular manner and by walking while using their mobile phones. For texting while walking, the participants were instructed to engage in familiar activities such as checking the social media and messaging their friends.

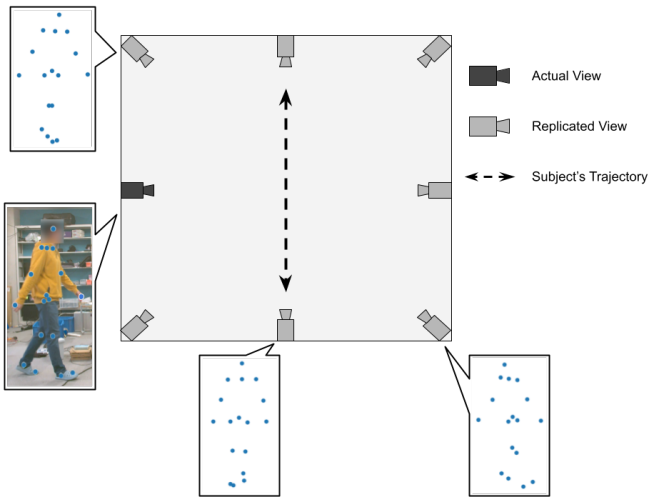

Fig. 4. Replicated Camera Perspectives.

The participants were recorded with 6 cameras arranged in the room, capturing a video from various angles. Subsequently, the videos were analyzed by Theia Markerless [18], a markerless motion capture system. The 3D world coordinates of the body key points (head, neck, shoulders, elbows, hands, pelvis, hip joints, knees, heels, and toes) of the participant were extracted from the videos.

One camera was located at a height of 1 meter from the floor to replicate the view akin to that of a camera mounted on a mobile robot. To facilitate the training and testing of the proposed method, the world coordinates of the body key points were translated into 2D camera coordinates. Pedestrian height was computed by taking the difference between the maximum and minimum y-coordinates of the body key points. The coordinates were normalized by adding a 15% margin above and below the body, creating a bounding box with a 1:2 width-height ratio, centered around the body key points' midpoint. The bounding box dimensions were then normalized to have a width and height of 1. Consequently, the normalized coordinates of the key points of pedestrian's body key points are obtained. As reported in [5], [6], it is deemed that the upper body of pedestrians contains features for identifying text walkers such as the position of the hands, angle of the elbows, and angle of the head. For this reason, the coordinates of the pedestrian's upper body key points (head, neck, shoulders, elbows, and hands) served as inputs during both training and testing phase,

Finally, to replicate the data viewed from cameras at different position and angles, the transfer and rotation matrices from world to camera coordinates were manipulated, such that the dataset contained data viewed from 360 degrees angle with an increment of 45 degrees, depicted in Fig. 4.

One trial of data is defined as the 2D coordinates of pedestrian upper body key points over 4 seconds taken at 50 FPS. The dataset comprises a total of 1,824 trials of data, representing 228 instances of texting while walking and normal walking viewed from 8 different angles.

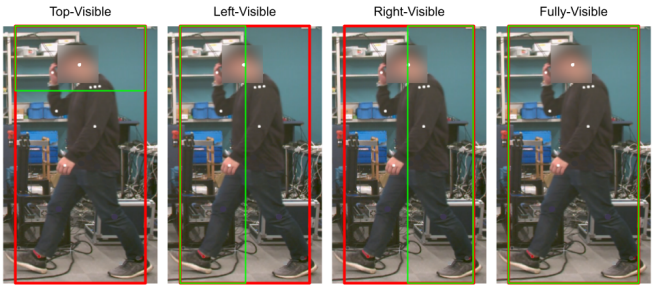

Fig. 5. Examples of Occlusion Patterns. Red box represents pedestrian's bounding box, and green box represents visible region.

For testing occluded pedestrians, the preprocessed data were further processed to replicate the data under occlusion. According to Dollar et al. [8], the majority of pedestrians are occluded from either below or the side, and 79% of occluded pedestrians are under heavy or full occlusions, meaning at least 40% of their body are occluded. To replicate such situation, four occlusion patterns were applied: top-visible, left-visible, right-visible, and fully-visible. Examples are shown in Fig. 5, where the red box depicts the bounding box and the green box depicts the visible region for the occlusion pattern. Top-visible is the case where only the top quarter of the normalized data is visible. Left-visible and right-visible are the cases where only the left and right half of the normalized data is visible, respectively. Finally, fully-visible is the case where all upper body key points remain visible. In addition, in all occlusion patterns, each body key point was randomly occluded with a probability of 10% following the report in [8].

B. Implementation of Supervisor Variational Autoencoder

The supervisor VAE is structured with a total of seven fully connected (FC) layers and a 2D latent space. The encoder network comprises two FC layers with 1024 units and rectified linear unit (ReLU). This is followed by two FC layers with 2 units with a linear activation function, generating μ and σ which are used to create a 2D latent vector z_{vae} via the reparametrization trick. The decoder network is consisted of two FC layers with 1024 units and ReLU, and another FC layer with 16 units and sigmoid activation function to reconstruct the input data from z_{vae}.

For training data, 70% of the data were randomly selected and the remaining were used as the test data. The model was trained for 50 epochs with Adam [19] as the optimizer. A KL coefficient β of 0.001 was chosen through experimentation, as it yielded the optimal result of having a low reconstruction loss while encoding the data of normal pedestrians and text walkers in distinct regions of the latent space.

C. Implementation of Text Walker Detection

The OIM consists of an LSTM layer with 16 units, an FC layer with 8 units and ReLU activation function, and an FC layer with 2 units with linear activation for generation z_{oim}. This is followed by the activity classification module which is an FC layer with 2 units with softmax activation

function. The hyperparameters were chosen via trial and error. A larger number of neurons may result in a better classification accuracy at an expense of computation cost, while too small number of neurons can lead to the opposite result.

The input to the OIM $O_{t-T:t}$ is 6 frames of body key points making up 1 second of observation of the pedestrian. This was done by downsampling the data collected, as sampling the data at any higher rate were considered to not provide much useful information since the movement of the body key points will be very small in between the adjacent frames. Furthermore, 1 second was chosen as it was considered that requiring a mobile robot to collect data for any longer time could increase the risk of collision with the pedestrian of interest.

Same as the implementation of supervisor VAE, 70% of data were used for training and the remaining were used for testing the model. The model was trained for 100 epochs with Adam as the optimizer. In our experiments, $w_{encoding}$ of 0.3 gave the optimal balance between $L_{encoding}$ and L_{cat}.

D. Ablation Test

The characteristic of the proposed method compared to that of previous research is that our method uses sequence of data and uses a pre-trained VAE to supervise the feature extraction process. To evaluate the contribution of the two factors, ablation test was conducted by training models that uses only single frame of data, and models trained without supervisor VAE. Our proposed model is named "6F with VAE" and the model trained with the same input but without the supervisor VAE is called "6F without VAE". Similarly, the model trained with single frame input and supervisor VAE is called "1F with VAE", and the single-input model trained without supervisor VAE is called "1F without VAE".

In addition, a CNN-based text-walker classification method proposed by Kumamoto et al. [5] was implemented to evaluate how our method performs in comparison to their method. We refer to this method as "Kumamoto et al."

V. RESULT

Figure 6 shows a plot of Receiver Operating Characteristic (ROC) of each model, showing the true and false positive rates at different threshold values of the classifier. The area under the ROC curve (AUC) was also calculated for each method to quantify their performance.

Our proposed method had the highest AUC of 0.945, followed by "6F without VAE" having an AUC of 0.943. In comparison, Kumamoto et al.'s method achieved an AUC of 0.908, "1F with VAE" had an AUC of 0.880, and "1F without VAE" had an AUC of 0.879.

The results underscore the impact of using sequential data as input, elevating the AUC by approximately 0.065. Furthermore, the incorporation of VAE contributed to a modest increase of 0.002, as evidenced by the comparison between models with and without VAE.

The F1 score was also calculated for each model as

$$F_1 = 2\frac{precision \cdot recall}{precision + recall}. \quad (5)$$

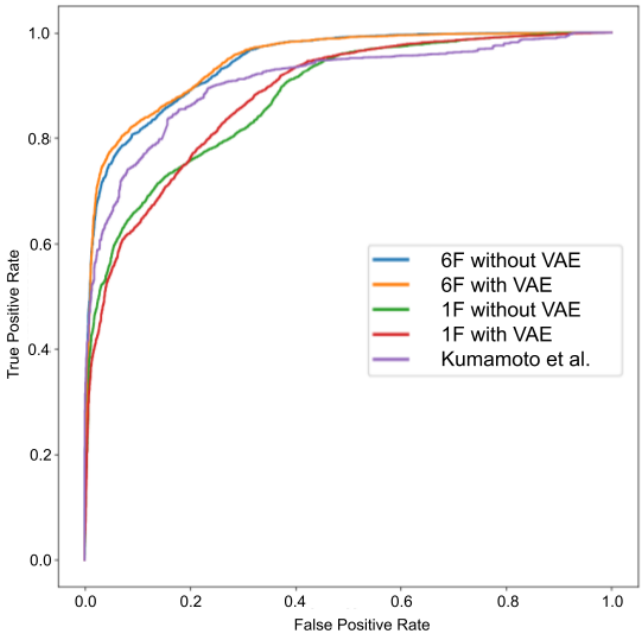

Fig. 6. ROC Curves of Different Methods.

TABLE I
CLASSIFICATION ACCURACY FOR DIFFERENT OCCLUSION PATTERNS.

Model	Occlusion Pattern			
	Top-Visible	Left-Visible	Right-Visible	Fully-Visible
6F with VAE	57.0	**91.6**	**90.4**	**94.4**
6F without VAE	49.3	90.6	89.7	94.3
1F with VAE	54.6	82.0	79.1	89.5
1F without VAE	51.7	76.9	77.0	90.1
Kumamoto et al.	**76.7**	76.8	82.6	92.2

The F1 score of the proposed method was 0.847, while that of the previous method was 0.817. The proposed method outperformed the previous method by 3%. The F1 scores of "6F without VAE", "1F without VAE", and "1F with VAE" were calculated to be 0.818, 0.798, and 0.757, respectively.

A. Occlusion Patterns

Table I summarizes the classification accuracy of each model across the four distinct occlusion patterns. The classification accuracy is calculated as the number of true positives and true negatives divided by the total number of data.

Leveraging multiple frames of data resulted in approximately 10% improvement for both left-visible and right-visible occlusion patterns. This enhancement can be attributed to the model's ability to predict body parts locations based on the movement of observed key points, even when only a portion of the pedestrian is visible. For example, if a pedestrian's hands are obscured, the model can infer their location by analyzing the movements of the elbows.

Integrating a supervisor VAE into the model contributes to a 6.7% and 2.9% increase in classification accuracy for top-visible data in multiple frames and single frame inputs, respectively. This outcome aligns with the hypothesis that a pre-trained VAE facilitates the extraction of meaningful data, even when dealing with limited data points.

Overall, the proposed method gave the best classification accuracy for three out of the four occlusion patterns. This underscores the pivotal role of utilizing sequential input data and a supervisor VAE in achieving robust results.

B. Pedestrian Behaviors

It was found that misclassifications are also caused by certain pedestrian behaviors resulting in a body pose similar to that of text walkers such as hands in front of the chest, bent elbows, and head facing downwards. For example, it was observed that the proposed model often misclassified a normal pedestrian as text walker when the pedestrian was touching their hair, or when the pedestrian was looking at the floor. The model was analyzed with SHapley Additive exPlanations [20] for these misclassified data to understand which features the model signified when making the classifications. It was found that the model often signified the coordinates of the hands, head, or neck, implying that one way the model distinguished between normal pedestrians and text walkers was based on if their hands were above their chests, or by checking if the pedestrian was facing downwards.

The findings suggest the proposed method's vulnerability against pedestrians with similar body pose as text walkers such as a person holding a cup of coffee.

VI. CONCLUSION

In this paper, we proposed an image-based machine learning method to discern text walkers from normal pedestrians, specifically focusing on scenarios involving occlusions. Unlike previous studies that relied on a single frame of data for text walker detection, our method addressed the challenge posed by occlusions: situations where feature extraction may falter due to limited data.

The proposed solution hinged on two key strategies: firstly, employing a pre-trained VAE to supervise the feature extraction process and secondly, leveraging a sequence of data as input. The model was trained and tested with body key points data obtained using Theia Markerless.

Results from the ablation test underscored the significance of the combined use of a supervisor VAE and sequential input data, showcasing an increased classification accuracy for three out of four occlusion patterns tested. Furthermore, the proposed method outperformed the previous state-of-the-art in terms of both AUC and F1 score. However, analysis revealed a potential area for improvements regarding the proposed method's susceptibility to pedestrians exhibiting a body pose similar to that of text walkers.

Future endeavors will involve real-life testing, employing a pose detector to extract body key points, instead of using an artificially occluded ground truth data. This approach may unveil new challenges such as instances where the pose detector misidentifies body parts of overlapping pedestrians. Additionally, an exploration of using the entire body's key points instead of just the upper body could open avenues to evaluate factors like pedestrian step sizes in classifying their activities.

REFERENCES

[1] L. L. Thompson, F. P. Rivara, R. C. Ayyagari, and B. E. Ebel, "Impact of social and technological distraction on pedestrian crossing behaviour: an observational study," *Injury Prevention*, vol. 19, no. 4, pp. 232–237, 2013.

[2] J. L. Nasar and D. Troyer, "Pedestrian injuries due to mobile phone use in public places," *Accident Analysis & Prevention*, vol. 57, pp. 91–95, 2013.

[3] F. Obayashi and K. Kozuka, "Sight property at the time 'texting while walking' by the gaze measurement, and its influence to walking," *IEICE Transactions*, vol. J100-A, no. 9, pp. 338–345, 2017.

[4] H. Murakami, C. Feliciani, Y. Nishiyama, and K. Nishinari, "Mutual anticipation can contribute to self-organization in human crowds," *Science Advances*, vol. 7, no. 12, p. eabe7758, 2021.

[5] K. Kumamoto and K. Yamada, "Detecting interaction of pedestrians with their smartphones based on body keypoints," in *Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 3261–3266.

[6] A. Rangesh and M. M. Trivedi, "When vehicles see pedestrians with phones: A multicue framework for recognizing phone-based activities of pedestrians," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 2, pp. 218–227, 2018.

[7] J. Wu, Y. Tamura, Y. Wang, H. Woo, A. Moro, A. Yamashita, and H. Asama, "Smartphone zombie detection from lidar point cloud for mobile robot safety," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2256–2263, 2020.

[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Proceedings of the International Conference on Learning Representations*, 2014.

[10] A. Shikishima, K. Nakamura, and T. Wada, "Proceedings of the detection of texting while walking by using smartphone's posture and acceleration information for safety of pedestrians," in *2018 16th International Conference on Intelligent Transportation Systems Telecommunications (ITST)*, 2018, pp. 1–6.

[11] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6995–7003.

[12] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000, pp. 417–424.

[13] H. Iwasaki, H. Hayashi, M. Kamezaki, and S. Sugano, "Driver pose estimation from steering-wheel occluded image by using image inpainting," *Transactions of Society of Automotive Engineers of Japan*, vol. 53, no. 3, 2023.

[14] O. Afolabi, K. Driggs-Campbell, R. Dong, M. J. Kochenderfer, and S. S. Sastry, "People as sensors: Imputing maps from human actions," in *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2342–2348.

[15] K. Hara, H. Kataoka, M. Inaba, K. Narioka, R. Hotta, and Y. Satoh, "Predicting appearance of vehicles from blind spots based on pedestrian behaviors at crossroads," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11 917–11 929, 2022.

[16] Y.-J. Mun, M. Itkina, S. Liu, and K. Driggs-Campbell, "Occlusion-aware crowd navigation using people as sensors," in *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 12 031–12 037.

[17] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[18] "Theia markerless," Aug 2022. [Online]. Available: https://www.theiamarkerless.ca/

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.