Video-based pedestrian crossing intention inference using contextual information for urban autonomous driving

Yujin Kim Yunyoung Kook Donggeon Lee Sunghee Park
{zxc123xc, chloecook0524, leedonggeon, adak0102}@snu.ac.kr
Seoul National University

Abstract

One of the significant challenges for autonomous vehicles in urban environments is comprehending and predicting other road users' actions, especially pedestrians at the point of crossing. The prevalent approach to solve this problem is to use the motion history of the pedestrian to predict their forthcoming trajectories or use the skeleton feature to inference intention. However, pedestrians exhibit extremely variable actions, most of which cannot be estimated without visual observation of the pedestrians and their surrounding road structure. Therefore, this paper introduces a vision-based model using contextual information for the pedestrian's intent classification problem at the point of crossing. The contextual information has been fully employed using the pixel-based raw image as input data. In this approach, the crossing intention of pedestrians has been binary classified through the pre-processing network and Resnet3D-based transfer learning.

1. Introduction

Over the last few decades, there has been a rapid growth in autonomous driving systems capable of performing various perception, planning, and control tasks. However, ensuring safety is one of the most challenging tasks for driving in highly dynamic urban environments. The safety of pedestrians has been significantly highlighted after the deadly crash in Arizona involving an autonomous vehicle operated by Uber [1]. Pedestrians are the most vulnerable road users and require an active protection system [2]. In terms of object detection, a considerable amount of research has already been done and applied practically to localize and classify pedestrians. R-CNN, Faster R-CNN, YOLO, and SSD are widely-used networks ensuring high accuracy and real-time application [3–6].

Nevertheless, for safe autonomous driving, not just the class of objects but the classification of pedestrian's intention can be required. Comprehending pedestrians' underlying intent and predicting their future actions in advance helps the driving systems to select the correct course of action to avoid any potential collisions and disruption of traffic flow [7]. This is particularly crucial when dealing with pedestrians at the point of crossing since they exhibit highly variable behavior patterns. However, it is pretty tricky task for autonomous system to interpret pedestrians' crossing intentions. As pedestrians are complex individuals, their intention to cross the street is affected by many factors, including contextual interaction between the surrounding traffic environment [8].

Suppose the autonomous vehicle in an urban environment plans its motion, knowing pedestrians' orientation and relative position from a vehicle. Three scenarios are illustrated in Figure 1. The distance between ego vehicle and pedestrian is similar in the A scenario and B scenario. However, the ego vehicle should decelerate or stop in case of A while it can keep driving in case B. The reason is that the pedestrian in scenario A is predicted to have crossing intent, whereas just walking along the sidewalk in scenario B. On the other hand, in scenario C, there is a crosswalk in front of the pedestrian. Thus, the ego vehicle can decelerate in scenario C even though the relative position between pedestrian and ego vehicle is the same in both cases. As demonstrated, contextual information such as pedestrian movements or positional relatives from road structures can be helpful to predict the crossing intent of pedestrians. This intent can be predicted from the pedestrian's movements or positional relatives from road structures based on contextual information from vision data.

This paper suggests a binary classifier that detects pedestrians intending to cross into an autonomous vehicle's lane based on spatial-temporal information from image sequences. The PIE Dataset [9] is used for training and testing, and Resnet3D with 18 layers, pre-trained on the Kinetics- 400 dataset, is chosen as the network backbone. The process can be distributed mainly into three steps. Firstly, based on the bounding box of the pedestrian, image pre-processing will be conducted to crop ROI from the entire image. Secondly, The cropped image is fused

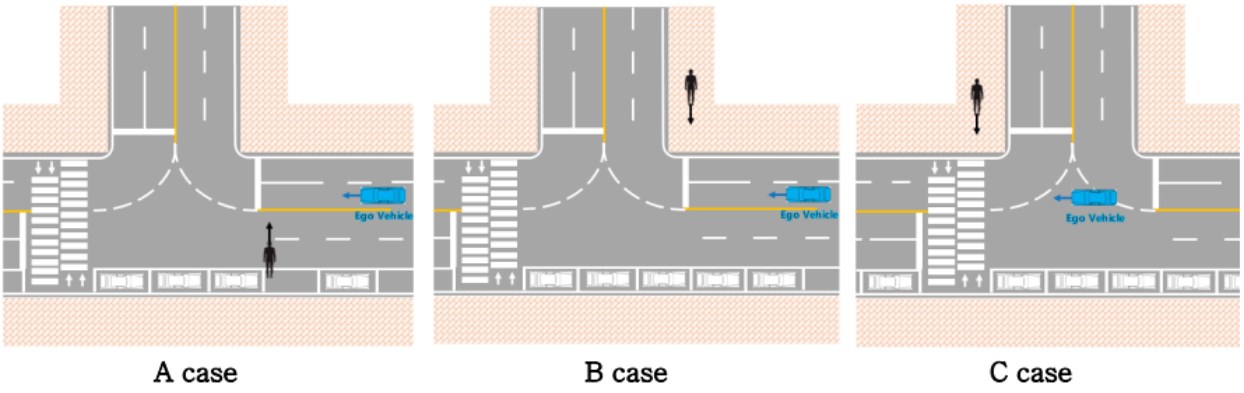


Figure 1. The diverse behavior scenarios of pedestrians

on the pixel level with a drivable area segmentation image through the YOLOP network. Thirdly, transfer learning is conducted using Resnet3D, and the intention of the target pedestrian will be classified. The performance of the proposed algorithm is verified based on PIE large-scale dataset. Overall, the suggested model achieves 73% accuracy and a 0.74 F1 score, showing stable accuracy on both "not-crossing" and "crossing" cases. The results prove that the presented methodology is effective for identifying pedestrian intentions on autonomous driving.

2. Related works

As the pedestrian is a complex individual, their intention can be influenced by many factors such as weather, the surrounding traffic environment, and even their own emotions [8]. This is why forecasting pedestrians' intentions sufficiently in advance is one of the most challenging tasks. However, the autonomous driving system must guarantee pedestrians' safety by proactive motion planning. In order to beat this problem, various approaches have been attempted to predict pedestrians' future behavior. The two major approaches are trajectory-based path prediction and intention inference using skeleton features, and context-based intention inferences are also getting attention these days.

2.1. Trajectory-based path prediction

The path prediction domain focus on the past observed trajectory of the pedestrians to predict future locations of pedestrians [9]. Most of the work in this approach is dedicated to predicting surveillance sequences where the movements of pedestrians are observed from a fixed bird's eye view perspective [10–18]. In a trajectory-based framework, early strategies for pedestrian detection and tracking used Kalman Filters [19], including interacting multiple model filters [20, 21], to account for different motion dynamics. Even so, the sole consideration of trajectory is deficient for accurately predicting the pedestrian path as the motion dynamics keep changing. Empirical studies have confirmed that a higher error rate is produced in drivers' judgment regarding the pedestrian intentions when only the pedestrian's trajectory is available [22]. Since they react to action already in progress instead of anticipating it, they are short-term predictions that are only effective when the pedestrians are already crossing or about to do so [9]. For instance, for trajectory-based approaches, a pedestrian walking alongside the road prior to the crossing or standing at the intersection can be challenging. Furthermore, the past trajectory of a pedestrian might not necessarily reflect their ultimate objective. A pedestrian waiting at a bus stop might step on the road to check for the bus, which can be interpreted as a crossing event by a trajectory-based approach. Additionally, lots of trajectory-based approaches are hard to apply with the camera directly mounted on a vehicle since they primarily rely on the bird-eye view camera.

2.2. Intention inference using skeleton feature

The other strategy is the skeleton-based intent prediction of pedestrians. These approaches are usually based on hand gestures, head orientation, and the body posture of the pedestrian [23]. E. Insafutdinov et al. [24] developed a tracking algorithm that simplified the body-part relationship graph and applied a feed-forward convolutional architecture to associate parts even in clutter. In another study, an algorithm called PoseTrack [25] was proposed. A graph with both spatial and temporal edges for detection is built. Then it simultaneously associates body parts within every single frame and each person over different frames by integer linear programming. However, these methods are designed for more general-purpose pose tracking and are still inefficient for pedestrian and cyclist tracking and pose recognition. Additionally, those approaches are mostly only available for short distances since high-resolution images are necessary to detect skeletons.

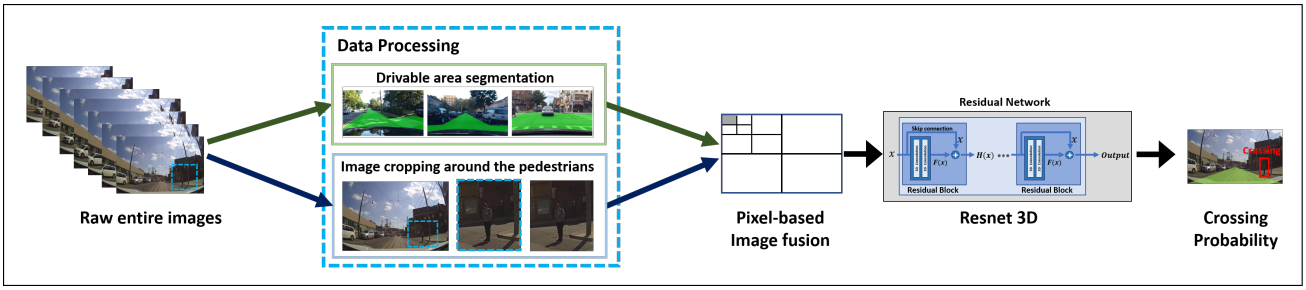


Figure 2. The architecture of the network

2.3. Context-based intention inference

As mentioned above, previous methods to predict pedestrians' intent have limitations. Some researchers have recently focused on contextual information as a remedy for the common drawbacks of trajectory-based and skeleton-based algorithms. Schneemann et al. [26] generated a feature vector by combining occupancy map of crosswalk/waiting area and bird's-eye-view history image of pedestrian from the road edge solved by scene segmentation. Then, classify the intention of the pedestrian by SVM classifier using this feature vector as input. However, additional processes are needed to apply this approach practically, such as the scene segmentation process, including human body orientation estimation, lane detection, and crosswalk/waiting area detection. Besides, contextual information initially contained in the raw image can be missed because they only take specific features from the raw image. Yang et al. [27] created a feature vector regarding the 2D pixel coordinates of pedestrians, the velocity of the on-coming vehicle, the existence of crosswalks, and traffic signals. Though, the research only considers the standing pedestrian and does not take into account spatial-temporal correlation information between pedestrian maneuver and road information. Rasouli et al. [9] provide a large-scale dataset for pedestrian intention estimation and suggest a convolutional LSTM based method using the dataset. Compared to previous researches, the study has practical strength on less pre-processing load as only the bounding boxes of detected pedestrians are used. Thus, using Rasouli et al.'s method [9] as baseline, this paper will propose improved model minimizing the loss of spatial-temporal correlation between road structure and pedestrian maneuver.

2.4. Datasets

A number of datasets for trajectory prediction contain videos collected from a perspective of top-down view [28–31] or surveillance camera perspective [32–34]. There are comparatively fewer datasets that are specifically provided for pedestrian behavior prediction from a moving vehicle perspective. Even though the publicly available pedestrian detection datasets [35–37] can potentially be used for such a purpose, they lack necessary characteristics such as ego-vehicle information [35], temporal correspondence [37], or enough pedestrians samples with long tracks [36]. These datasets also do not contain any form of pedestrian behavior annotations that can be employed for action prediction.

JAAD [38] is a newly introduced dataset that contains a large number of pedestrian samples with temporal correspondence, a subset of which are annotated with behavior information. However, for the aim of intention estimation and trajectory prediction, this dataset has several drawbacks. The dataset does not have ego-vehicle data, the videos are distributed into short discontinuous chunks, and most pedestrian samples with behavioral annotations have crossing intent.

As a primary dataset for this paper, the Pedestrian Intention Estimation (PIE) dataset [9] was chosen, consisting of 6 hours of driving footage in urban environments. The dataset provides bounding box annotations for pedestrians and traffic objects as well as sensor readings of the ego-vehicle and ego-motion data recorded from the camera.

3. Method

To understand the pedestrian's crossing intention, we address the problem based on local surroundings and motion of target pedestrians in consecutive frames. In this section, a detailed network structure and methodology are introduced to deal with this.

3.1. Network Architecture

The proposed algorithm is a binary classifier of whether pedestrians have the intention to enter the subject vehicle's driving lane. The probability of pedestrian crossing is derived as a final output through the proposed algorithm. The cumulative frame of the pixel-based image cropped the region of interest(ROI) from the entire image is used as an input. The ROI is set to an enlarged bounding box, including surrounding information of the target pedestrian. Contextual information around the target pedestrian can be used

without a loss since the raw image frames are used as input instead of the processed map from the raw image. The process for estimating pedestrian crossing intention is largely composed of two steps: drivable area fused image extraction and Resnet3D-based transfer learning. The overall architecture is illustrated in Figure 2.

3.2. PIE Dataset

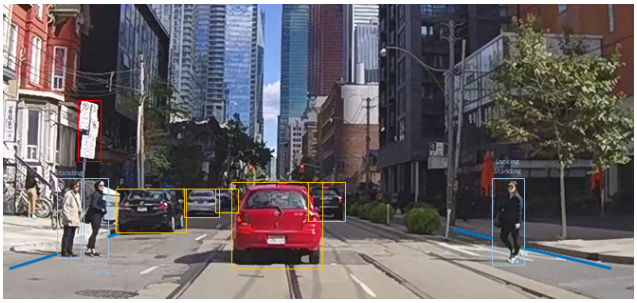
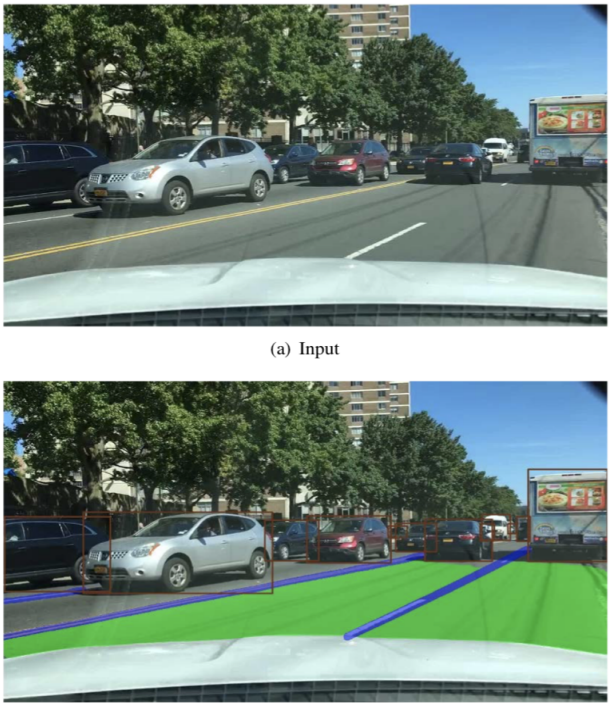


Figure 3. Example of annotations given in PIE dataset [9]

This study trains and tests the proposed model on the PIE dataset [9] which is a new dataset for studying pedestrian behavior in traffic. PIE contains 6 hours of HD video are recorded with an on-board camera at 30 FPS and split into approximately 10 minute chunks grouped into six sets. Bounding boxes are provided for 1842 pedestrians and vehicles that interact with the driver, as well as for elements of infrastructure such as traffic lights, signs, zebra crossings, road boundaries. Additionally, accurate ego-vehicle information from the OBD sensor is available synchronized with video footage. It includes speed, GPS coordinates, and heading direction. All videos were recorded in HD format (1920 × 1080 px) at 30 fps, split into approximately 10 minute long chunks, and grouped into six sets. The dataset represents a wide diversity of pedestrian behaviors at the crossing point, including the busy one-way street and wide boulevards with fewer pedestrians. PIE provides long continuous sequences and annotations for a wide range of applications. Rich spatial and behavioral tags are available for each pedestrian per frame as shown in Figure 3, including actions such as walking, standing, crossing, looking. Over 300K labeled video frames with 1842 pedestrian samples make PIE the largest publicly available dataset for studying pedestrian behavior in traffic.

3.3. Drivable area fused image extration

In order to reflect the behavior of pedestrians on the surrounding contextual information such as roads or walks, drivable area segmentation information is fused to the cropped raw image. The drivable area is obtained through YOLOP network [39] as presented in Figure 4. The fusion is processed through sum operation in pixel unit so that drivable area can be applied while maintaining RGB of the raw



(a) Input

(b) Output

Figure 4. Drivable area segmentation obatained from YOLOP [39]

image without information loss. That is, the drivable area is processed as an image with increased brightness without modifying the RGB ratio. The fused image then be an input to the main network, Resnet3D. This process helps model train the correlation between pedestrian movement and spatial information in the Resnet3D network by fusing road information which is hard to obtain from just cropped images.

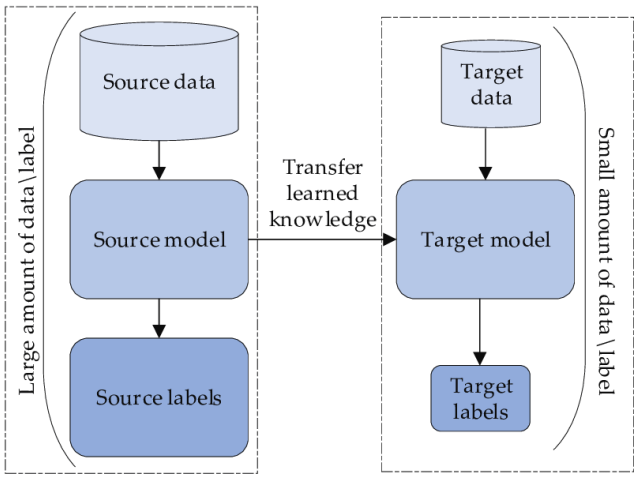


Figure 5. Concept of transfer learning [40]

	Method	Drivable area	Unfrozen block	Accuracy	F1 Score	Negative Accuracy	Positive Accuracy
experiment (a)	R3D		FC	0.58	0.55	0.35	0.8
	R(2+1)D		FC	0.61	0.59	0.38	0.84
	MC3		FC	0.59	0.56	0.31	0.88
experiment (b)	R3D		FC	0.58	0.55	0.35	0.8
	R3D		4 + FC	0.65	0.64	0.44	0.87
	R3D		3, 4 + FC	0.67	0.66	0.48	0.87
	R3D		2, 3, 4 + FC	0.68	0.67	0.54	0.82
	R3D		1, 2, 3, 4 + FC	0.69	0.69	0.63	0.76
	R3D		All	0.69	0.68	0.56	0.81
experiment (c)	R3D		1, 2, 3, 4 + FC	0.69	0.69	0.63	0.76
	R(2+1)D		1, 2, 3, 4 + FC	0.68	0.67	0.5	0.85
	MC3		1, 2, 3, 4 + FC	0.66	0.64	0.42	0.9
experiment (d)	R3D	✓	FC	0.52	0.52	0.52	0.52
	R3D	✓	4 + FC	0.73	0.74	0.68	0.78
	R3D	✓	3, 4 + FC	0.66	0.66	0.62	0.70
	R3D	✓	2, 3, 4 + FC	0.67	0.67	0.74	0.60
	R3D	✓	1, 2, 3, 4 + FC	0.55	0.54	0.58	0.51

Table 1. Main results

3.4. Resnet3D-based transfer learning

The following step is for learning spatial-temporal information from successive feature maps. We employed a ResNet3D [41] with 18 layers, pre-trained on the Kinetics-400 dataset [42]. As this source model was originally trained on the larger dataset for action recognition of video, we expected that the transfer learning might significantly boost our crossing intention inference model's performance as depicted in Figure 5. The transfer learning is performed by transforming and re-learning the top layer in order to infer pedestrian crossing intention. So, the representation is re-produced with the pre-trained model. Through this process, the correlation between the processed feature maps is reflected, and high-level features that represent the position on the surrounding context and the pose change of pedestrians are learned.

4. Experiments

In this section, our method is evaluated and validated. We first describe the implementation details, followed by experiment results and comparison to the previous state-of-the-art approach.

4.1. Implementation details

We conducted all experiments using RTX2070 super in the PyTorch framework. The cropped images fused with the driving area were resized to 112×112 pixels. Among the annotations provided by the PIE database, pedestrian ID, bounding box information, and crossing intention probability were used, and the probability was reconstructed into binary intention based on 0.5. One input sequence consisted of 15 frames, which is about 0.5 seconds, and the sequence overlap rate was used to increase the amount of learning data. Since PIE data has more positive situations than negative, the total number of data sets was balanced. The total number of data is 10696, and the ratio of trains, verification, and test is 0.55, 0.15, and 0.3, respectively, with 6068, 1613, and 3015. Adam, in which Adgrad and RMSPropoptimizer are fused, was used as an optimizer. Adam optimizer used an initial learning rate of 0.0001 and reduced it to 0.4 for every 3 epochs. In addition, if the valuation loss did not decrease during 5 epochs, early-stop was performed. Also, binary cross-entropy was used as the loss function. In this environment, we conducted various experiments with Resnet models, unfrozen block, and driving area and performances such as total accuracy, F1 score, negative and positive accuracy were derived.

4.2. Results

3D Pre-trained models. Experiment (a) in Table 1 shows the pre-trained model's performance with different networks. In this experiment, for video representation, Resnet3D(R3D), R(2+1)D, and mixed convolution 3 (MC3) networks have been tested. The main differences between the three models are the various filters used in each layer. The R3D uses convolutional filters of equal size in three dimensions in width, height, and time. The R(2+1)D model factorizes 3D convolutional filters into separate spatial and temporal components with different spatial and temporal sizes. And its total number of parameters is similar to the R3D model. The mixed convolution 3 (MC3) model combines 3D convolutional filters in the first nine layers with subsequent layers using 2D filters. In consideration of

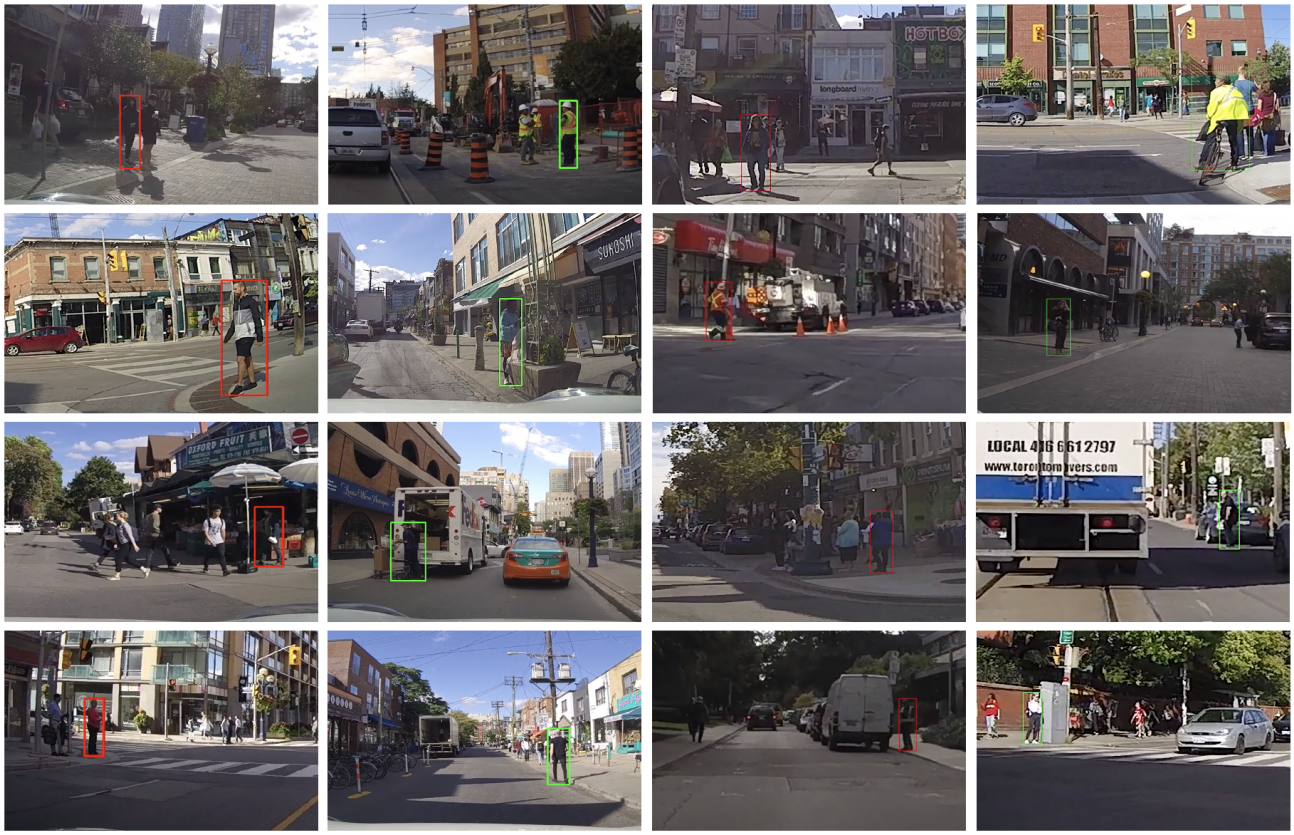


Figure 6. Visualized results of our pedestrian intention estimation overlaid on top of image frames from the PIE dataset (cropped for better visibility). Bounding boxes are colored depending on the presence (red) or absence (green) of crossing intention as detected by our model. Thin lines of bounding boxes represent incorrectly estimated intention.

accuracy and F1 score, the performance of R(2+1)D was highest, followed by MC3 and R3D when the unfrozen block is the same as the final layer(FC). However, when we only modified the final classification layer while leaving pre-trained weight, binary classification's accuracy for the negative case was poor compared to the positive case. Negative accuracy indicates when the model identifies pedestrians' intention not to cross, while positive accuracy only counts pedestrians with crossing intention. Through the experiment (a) given in Table 1, we observed that the negative accuracy was relatively low than the positive accuracy. For all three cases, the negative accuracy was low in the range of 0.3 0.4, which means the model easily misjudges not-crossing intention to crossing intention when most layers are frozen. Therefore, we tried to modify the number of unfrozen blocks to improve negative accuracy.

Unfreeze. In order to train the model to learn video representation, additional experiments (b) were conducted by unfreezing previous layers, and the result is given in 1. As the number of unfrozen blocks increases, the model's performance tends to improve. On the other hand, accuracy drops when we unfreeze the stem layer extracting features through CNN. Hence we let the stem layer frozen since the performance was better when the pixel-level(low-level)representation was extracted from the stem layer with pre-trained model's weight. The highest accuracy was achieved when we trained the model with high-level representation and left the stem layer frozen. Negative accuracy increases to 0.63 as the number of learning layers increases, leading to higher overall accuracy. As shown in experiment (c) of Table 1, the performance was best in the R3D model with 0.69 of accuracy when every layer was unfrozen except the stem layer. R(2+1)D and MC3 networks followed next with the accuracy of 0.68 and 0.66.

Drivable area. Even though the prior experiments confirmed that R3D networks with unfreezing can improve accuracy, the negative accuracy was still insufficient as 0.63. To solve the problem, the drivable area segmentation was added at the data processing stage. The R3D model's performance with the drivable area is given in experiment (d) of Table 1. Compared to the experiment (b), the input data fused with drivable area segmentation helped the model get

improved negative accuracy.

Combination of Unfreeze and Drivable area. We also examined how the combination of unfreezing and drivable area impact the performance of the model(see experiment (d) in Table 1). The R3D model with drivable area showed the highest accuracy when the 4th block and last layer(FC) were unfrozen. With this configuration, the negative accuracy reached 0.68, and total accuracy was 0.73, which is state-of-the-art. Therefore, we accepted this configuration as our final model and the visualized output of our model is given in Figure 6.

Previous intention reflection. In the corresponding structure, the input sequences are divided by the set number of input frames even for the same person and, the correlation between the divided sequences for the same person is not considered. Therefore, We attempted to reflect the previously estimated intention for the same pedestrian ID sequence to the current intention estimation. The attempt was implemented in two ways. The first is to fuse the current output and the previous output using the concept of momentum. In the second method, a new loss term was added to follow the average value of the previous stacked output. However, both methods resulted in poor performance. In the 3D CNN structure, back propagation is not possible from the previous output value, so it is estimated that the term added as momentum acts as a variable bias, which actually degrades the performance. In addition, in the case of adding loss term, if the judgment from the first sequence input is incorrect, the performance of the subsequent output may be adversely affected. Therefore, the reflection of previous intention is not applied to our final model.

4.3. Comparison to the previous state-of-the-art

Method	Balancing	Additional Info	Acc.	F1 Score
Our model	No	No	0.78	0.79
Baseline(PIE)	No	No	0.69	0.79
Baseline(PIE)	No	Yes	0.79	0.87
Our model	Yes	No	0.69	0.69
Our model	Yes	Yes	0.73	0.74

Table 2. comparison between our model and PIE [9]'s model

The model and performance proposed by PIE [9] talk about the results for unbalanced data. In this case, the data with labels of positive (trying to cross) and negative (not to cross) are unbalanced to about 4.5:1.

In the case of unbalanced data, Our baseline model outperforms PIE's model without using additional information(see Table 2. When PIE's model uses additional information such as context, bounding boxes, and bounding boxes coordinates, its accuracy is 0.79, and the F1 score is 0.87. With this unbalanced data, our baseline without additional information shows the performance of 0.77 accuracy and 0.78 F1 score.

In an environment that uses unbalanced and biased data as input, the PIE model that easily inference intentions to crossing case shows a performance of 0.82. Therefore, it does not seem appropriate to be used for performance judgment.

Even though there are no results for the PIE model predicted based on the balancing data, we wanted to compare further our baseline and our SOTA based on the balancing data. In comparison, the performance for the negative was significantly improved, and our SOTA outperformed our baseline model showing great advancement.

5. Conclusion

In this paper, we proposed a model to classify pedestrians' crossing intention. By evaluating various networks and combinations, we showed that Resnet-3d based transfer learning from the action recognition model is a good predictor for crossing intention. Our baseline model outperforms the previous state-of-the-art method without any additional input data but raw images. Also, We prove that drivable area segmentation can further improve the performance of the model. Our best model with segmentation shows improvement on negative accuracy when tested on balanced input data, while the previous method only tested on the biased dataset. Overall, the presented model is conceptually and computationally simpler than the previous methods, ensuring reliable accuracy for identifying both not-crossing and crossing intention.

For future work, the proposed model can be tested on other datasets such as JAAD, and real-time applications also can be validated by implementation on autonomous vehicles. Additionally, the crossing intention estimation might be further enhanced by appropriately considering pedestrians' previous intentions and social interactions, all of which affect future pedestrian actions.

References

[1] P. Kohli and A. Chadha, "Enabling pedestrian safety using computer vision techniques: A case study of the 2018 uber inc. self-driving car crash," in *Future of Information and Communication Conference*, pp. 261–279, Springer, 2019. 1

[2] C. Hilario, J. M. Collado, J. M. Armingol, and A. de la Escalera, "Pedestrian detection for intelligent vehicles based on active contour models and stereo vision," in *International Conference on Computer Aided Systems Theory*, pp. 537–542, Springer, 2005. 1

[3] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2017. 1

[4] H. Zhang, Y. Du, S. Ning, Y. Zhang, S. Yang, and C. Du, "Pedestrian detection method based on faster r-cnn," in

2017 13th International Conference on Computational Intelligence and Security (CIS), pp. 427–430, IEEE, 2017. 1

[5] W. Lan, J. Dang, Y. Wang, and S. Wang, "Pedestrian detection based on yolo network model," in *2018 IEEE international conference on mechatronics and automation (ICMA)*, pp. 1547–1551, IEEE, 2018. 1

[6] D. Liu, S. Gao, W. Chi, and D. Fan, "Pedestrian detection algorithm based on improved ssd," *International Journal of Computer Applications in Technology*, vol. 65, no. 1, pp. 25–35, 2021. 1

[7] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked rnns," *arXiv preprint arXiv:2005.06582*, 2020. 1

[8] S. Ji, Y. Peng, H. Zhang, and S. Wu, "An online semisupervised learning model for pedestrians' crossing intention recognition of connected autonomous vehicle based on mobile edge computing applications," *Wireless Communications and Mobile Computing*, vol. 2021, 2021. 1, 2

[9] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6262–6271, 2019. 1, 2, 3, 4, 7

[10] Y. Hu, S. Chen, Y. Zhang, and X. Gu, "Collaborative motion prediction via neural motion message passing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6319–6328, 2020. 2

[11] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14424–14432, 2020. 2

[12] H. Sun, Z. Zhao, and Z. He, "Reciprocal learning networks for human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7416–7425, 2020. 2

[13] J. Sun, Q. Jiang, and C. Lu, "Recursive social behavior graph for trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 660–669, 2020. 2

[14] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *European Conference on Computer Vision*, pp. 759–776, Springer, 2020. 2

[15] C. Choi and B. Dariush, "Looking to relations for future trajectory forecast," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 921–930, 2019. 2

[16] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12085–12094, 2019. 2

[17] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1349–1358, 2019. 2

[18] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264, 2018. 2

[19] D. Llorca, M. Sotelo, A. Hellín, A. Orellana, M. Gavilán, I. Daza, and A. Lorente, "Stereo regions-of-interest selection for pedestrian protection: A survey," *Transportation research part C: emerging technologies*, vol. 25, pp. 226–237, 2012. 2

[20] M. E. Farmer, R.-L. Hsu, and A. K. Jain, "Interacting multiple model (imm) kalman filters for robust high speed human motion tracking," in *Object recognition supported by user interaction for service robots*, vol. 2, pp. 20–23, IEEE, 2002. 2

[21] Y. Boers and J. N. Driessen, "Interacting multiple model particle filter," *IEE Proceedings-Radar, Sonar and Navigation*, vol. 150, no. 5, pp. 344–349, 2003. 2

[22] S. Schmidt and B. Faerber, "Pedestrians at the kerb–recognising the action intentions of humans," *Transportation research part F: traffic psychology and behaviour*, vol. 12, no. 4, pp. 300–310, 2009. 2

[23] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Understanding pedestrian behavior in complex traffic scenes," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 1, pp. 61–70, 2017. 2

[24] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "Arttrack: Articulated multi-person tracking in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6457–6465, 2017. 2

[25] U. Iqbal, A. Milan, and J. Gall, "Posetrack: Joint multi-person pose estimation and tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2020, 2017. 2

[26] F. Schneemann and P. Heinemann, "Context-based detection of pedestrian crossing intention for autonomous driving in urban environments," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 2243–2248, IEEE, 2016. 3

[27] B. Yang, W. Zhan, P. Wang, C. Chan, Y. Cai, and N. Wang, "Crossing or not? context-based recognition of pedestrian crossing intention in the urban environment," *IEEE Transactions on Intelligent Transportation Systems*, 2021. 3

[28] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer graphics forum*, vol. 26, pp. 655–664, Wiley Online Library, 2007. 3

[29] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 261–268, IEEE, 2009. 3

[30] B. Majecka, "Statistical models of pedestrian behaviour in the forum," *Master's thesis, School of Informatics, University of Edinburgh*, 2009. 3

[31] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European conference on computer vision*, pp. 549–565, Springer, 2016. 3

[32] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR 2011*, pp. 3153–3160, IEEE, 2011. 3

[33] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *CVPR 2011*, pp. 3457–3464, IEEE, 2011. 3

[34] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2871–2878, IEEE, 2012. 3

[35] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 304–311, IEEE, 2009. 3

[36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361, IEEE, 2012. 3

[37] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3221, 2017. 3

[38] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 206–213, 2017. 3

[39] D. Wu, M. Liao, W. Zhang, and X. Wang, "Yolop: You only look once for panoptic driving perception," 2021. 4

[40] L. Alzubaidi, M. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, Y. Duan, and S. Oleiwi, "Towards a better understanding of transfer learning for medical imaging: A case study," *Applied Sciences*, vol. 10, p. 4523, 06 2020. 4

[41] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," 2018. 5

[42] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017. 5