Pedestrian's Intention Recognition, Fusion of Handcrafted Features in a Deep Learning Approach

Omar Hamed, H. Joe Steinhauer
University of Skövde
Högskolevägen 1, 54128 Skövde
omar-hamed@hotmail.com, joe.steinhauer@his.se

Abstract

The safety of vulnerable road users (VRU) is a major concern for both advanced driver assistance systems (ADAS) and autonomous vehicle manufacturers. To guarantee people safety on roads, autonomous vehicles must be able to detect the presence of pedestrians, track them, and predict their intention to cross the road. Most of the earlier work on pedestrian intention recognition focused on using either handcrafted features or an end-to-end deep learning approach. In this project, we investigate the impact of fusing handcrafted features with auto learned features by using a two-stream neural network architecture. Our results show that the combined approach improves the performance. Furthermore, the proposed method achieved very good results on the JAAD dataset. Depending on whether we considered the immediate frames before the crossing or only half a second before that point, we received prediction accuracy of 91%, and 84%, respectively.

Introduction

According to the European Commission, pedestrians and cyclists were involved in 29% of fatalities in 2018 in the EU[1]. In order to ensure pedestrian's safety in the future, incorporating pedestrian intention recognition within advanced driver assistance systems (ADAS) and autonomous vehicles is crucial. Intention recognition is defined as the problem of inferring the intention (high-level goal) of an agent through its actions and the effects of these actions on the environment (Han and Pereira 2013). Automating this intention recognition is not a trivial task as the pedestrian's intention to cross does depend on a wide range of factors (Varytimidis et al. 2018).

The majority of earlier work in this area focused on using handcrafted features (manually designed or defined by a subject-matter expert) such as the pedestrian's head orientation, motion, distance to the curb, etc. to predict pedestrians' intention (Varytimidis et al. 2018; Fang and Lopez 2019). Some recent studies, however, addressed this problem by using an end-to-end deep learning approach (where the whole solution from sensor input to the final result or prediction is managed through a neural network) (Ghori et al. 2018; Saleh, Hossny, and Nahavandi 2019). In this work we propose fusing handcrafted features such as the pedestrians head orientation and automatically learned features by using a two-stream neural network architecture.

[1] ec.europa.eu/commission/presscorner/detail/en/qanda_20_1004

Features Fusion Approach

Deep learning methods have been able to automatically recognize and learn a wide range of low-level features and were able to outperform traditional handcrafted methods in different domains. Our hypothesis is that fusion of handcrafted features and learned features can add additional information, enforce the learning of these features, and hence improve the model performance.

To validate this hypothesis, we first created a baseline model consisting of a convolutional neural network to learn the visual representation, an LSTM network responsible for learning the temporal context, followed by a dense network that works as a classifier of the pedestrian intention. The baseline model was trained and tuned using only the pedestrian cropped images. For feature fusion, we propose instead a new two-stream architecture, as shown in Figure 1.

In the first stream, we use the pre-trained baseline model excluding the dense network layers, and in the second stream we apply an LSTM network to learn the temporal context of the handcrafted features. After concatenating these two streams, their output is fed into a dense network that acts as a classifier to predict the pedestrian intention.

We carried out two sets of experiments using the Joint Attention in Autonomous Driving (JAAD) dataset (Rasouli, Kotseruba, and Tsotsos 2017). In the first set we considered the latest 8 frames before the pedestrian's actual action (crossing or not) while in the second set we tried to predict the pedestrian intention 0.5 seconds ahead of the actual

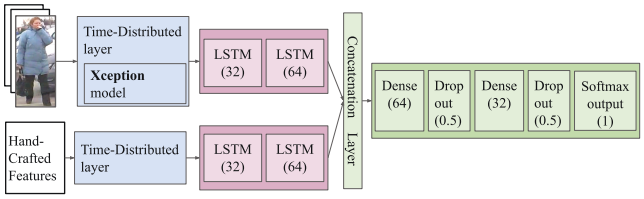

Figure 1: Proposed Feature Fusion Model Architecture.

action. This was achieved by considering the 8 frames that come before the latest look-ahead 16 frames as illustrated in Figure 2.

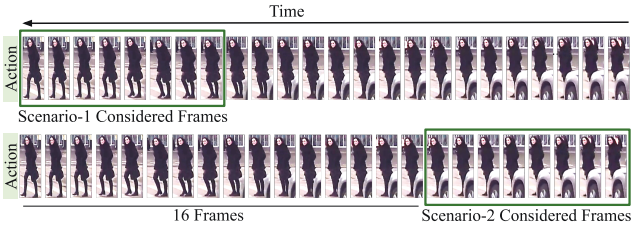

Figure 2: Considered Frames in Both Scenarios.

In our fusion experiments, we considered a list of handcrafted pedestrian-related features such as head orientation, hand gesture, and nodding all features annotated in the JAAD dataset. Additionally, we considered the two calculated features of the pedestrian's distance to the car and the pedestrian's distance to the curb as well as the vehicle acceleration behavior which also is available in the JAAD dataset.

Results, Discussion and Future Work

The dataset was divided into 70% for training, 15% for validation, and 15% for testing. We evaluated our models on the testing dataset and reported both classification accuracy and F1 score measures. Table 1 shows the evaluation results of our models on the testing dataset alongside the reported results of (Fang and Lopez 2019) and (Ghori et al. 2018). However, as the datasets used were different in all three approaches, these results cannot be directly compared. Nevertheless, for the first scenario where no look-ahead frames were considered, a classification accuracy of 91% and F1-score of 90% was achieved. This gives rise to form the hypothesis that our fusion model can potentially outperform the other models.

Our results show further that the fusion of handcrafted features has significantly higher accuracy than our own baseline model, especially in the first scenario where no look-ahead was used. It appears that the fusion of the head orientation feature helped to increase the prediction accuracy by 7 percentage points. This gives rise to the hypothesis that the use of additional handcrafted features along with the deep learned features could potentially further improve the performance of a pedestrian intention recognition solution. Investigating the impact of other handcrafted features and replicating the results on other datasets are hence important parts in our future work.

Approach	**Dataset**	**Acc**	**F1 Score**
Handcrafted features based on 2D pose (Fang and Lopez 2019)	JAAD	88	-
Deep learning 2D pose estimation key points (Ghori et al. 2018)	Daimler, Hanau, YouTube	-	87
Our baseline (only images)	JAAD	84	82
Our fusion model (Head Orientation)	JAAD	91	90
Our baseline model (16 look-ahead frames)	JAAD	82	78
Our fusion model (16 look-ahead frames)	JAAD	84	81

Table 1: Our Results and Recent State-of-The-Art Results

Acknowledgments

This work has been supported by VINNOVA, the Swedish Government Agency for Innovation Systems, proj. "Intention Recognition for Real-time Automotive 3D situation awareness (IRRA)", in the funding program FFI: Strategic Vehicle Research and Innovation (DNR 2018-05012).

References

Fang, Z.; and Lopez, A. M. 2019. Intention Recognition of Pedestrians and Cyclists by 2D Pose Estimation. *IEEE Trans. Intell. Transport. Syst.* 1–1. doi:10.1109/TITS.2019.2946642.

Ghori, O.; Mackowiak, R.; Bautista, M.; Beuter, N.; Drumond, L.; Diego, F.; and Ommer, B. 2018. Learning to Forecast Pedestrian Intention from Pose Dynamics. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, 1277–1284. Changshu: IEEE. doi:10.1109/IVS.2018.8500657.

Han, T. A.; and Pereira, L. M. 2013. State-of-the-art of intention recognition and its use in decision making. *AI Communications* 26(2): 237–246. doi:10.3233/AIC-130559.

Rasouli, A.; Kotseruba, I.; and Tsotsos, J. K. 2017. Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 206–213. Venice, Italy: IEEE. doi:10.1109/ICCVW.2017.33.

Saleh, K.; Hossny, M.; and Nahavandi, S. 2019. Real-time Intent Prediction of Pedestrians for Autonomous Ground Vehicles via Spatio-Temporal DenseNet. In *2019 International Conference on Robotics and Automation (ICRA)*, 9704–9710. Montreal, QC, Canada: IEEE. doi:10.1109/ICRA.2019.8793991.

Varytimidis, D.; Alonso-Fernandez, F.; Duran, B.; and Englund, C. 2018. Action and intention recognition of pedestrians in urban traffic. *arXiv:1810.09805 [cs]* URL http://arxiv.org/abs/1810.09805. ArXiv: 1810.09805.