Towards Multi-Modal Crash Prediction Based on V2X and Visual Information Using a Social Robot

Ludivine Morales*, Manuel Bied* and Alexey Vinel*†
*Karlsruhe Institute of Technology, Karlsruhe, Germany
†Halmstad University, Halmstad, Sweden
Email: luthosu@gmail.com, {manuel.bied, alexey.vinel}@kit.edu

***Abstract*—The development of autonomous vehicles and vehicular communications (V2X) promises to significantly enhance traffic safety and efficiency. However, challenges remain in ensuring safe interactions between autonomous vehicles and vulnerable road users (VRUs). We promote the idea to use a social robot as interface between social interaction and V2X. We propose to use such a robot for crash prediction: the social robot, equipped with an RGB-D camera and V2X-communication capabilities, gathers data on pedestrians' trajectories and vehicles' movement within a shared environment. The data can then be used to predict possible crashes. In this ongoing work, we present a framework that integrates the basic functionality to implement such an approach. To test the approach a data set consisting of videos of crossing pedestrians and V2X data of an eBike was collected. The system effectively converts the trajectories of pedestrians and vehicles into a shared coordinate frame, enabling precise detection of potential collisions. The preliminary findings show potential for a novel method for crash prediction.**

***Index Terms*—V2X, Vulnerable Road Users, Pedestrians, Collective Perception, Autonomous Vehicles, Crash Prediction, Traffic Robot.**

I. Introduction

Vehicular communications (V2X) promises to significantly improve road safety and efficiency [1], [2]. However, how this technology can be beneficially used for Vulnerable Road Users (VRUs) is still an active research topic. A major advantage of the V2X-communication is the fact that it is an non-line of sight (NLOS) approach, i.e. it can also be used over large distances or when the direct line of sight is blocked. For VRUs, like pedestrians, to fully benefit from this technology, they must be in possession of V2X-enabled devices [3]. While smartphones present a potential solution, they may introduce distractions and might not be ideal for the use in traffic [4], [5]. Other disadvantages include privacy concerns and that not all VRUs necessarily carry smartphones with them.

The question of how to communicate with VRUs is further strengthened with autonomous vehicles (AV) on the horizon. While AVs promise various benefits (comfort, efficiency and safety), the lack of a human driver requires to develop alternative approaches for their communication with other road users [6]. To this end, we promote the idea of using a social-interactive robot to communicate with nearby vehicles through V2X-communication, monitor the environment with its own sensors and interact with road users (here more specifically VRUs) via natural communication (e.g. gestures and speech) or its display. In this sense the robot could act as mediator for the different actors in traffic. The advantages and use-cases of the robot are manyfold. One advantage lies in its ability to serve as an interface between autonomous agents using digital communications and human actors using social interaction. Additionally, a robot seems to be well suited to draw the attention of pedestrians [7]. The physical presence of the robot is likely to increase its persuasiveness [8]. Considering these points together, the robot offers a promising possibility to extend the advantages of V2X to road users that can not receive V2X communication themselves. Another application is to use the robot's sensor data to contribute to road safety. The robot can contribute by communicating its (processed as necessary) sensor data via V2X. While this can also be done by stationary infrastructure, the advantage of the robot is its mobility. It can, for example, move within the space of pedestrians and focus on areas with high uncertainty.

One promising application is the integration of V2X communication with robotic visual perception offering a new dimension to the prediction of accidents. The robot can simultaneously monitor pedestrians' trajectories using visual perception and vehicles' trajectories received via V2X. This

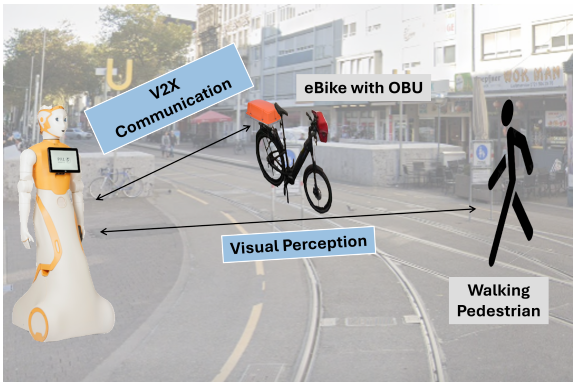

Fig. 1. Social robot collecting V2X and Visual Information in the test environment. The data is used to calculate the distance between pedestrian and eBike to predict a possible crash.

This work has been funded as part of the KIT Future Fields project "V2X4Robot". This paper is part of the CulturalRoad project, funded by the European Union under grant agreement No. 101147397. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

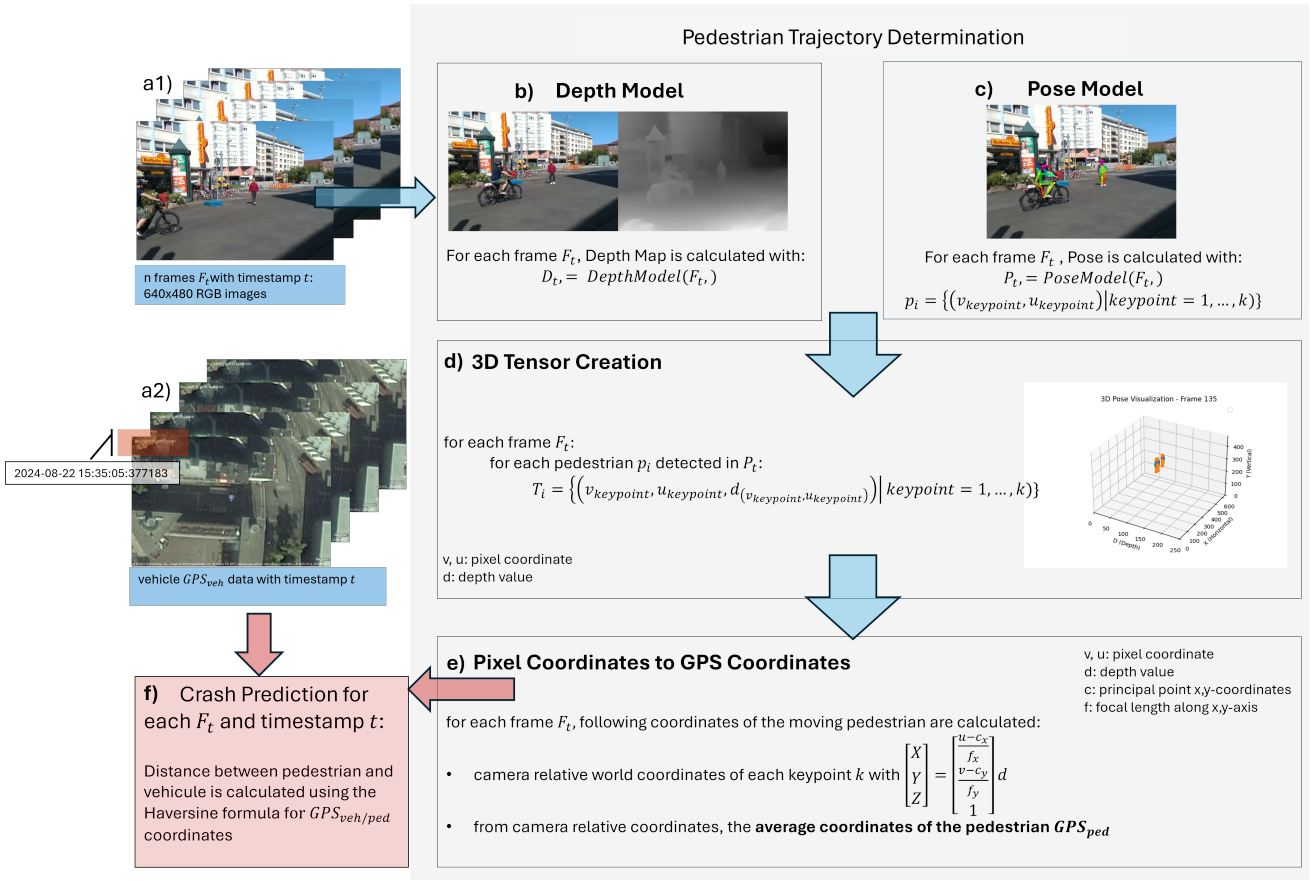

Fig. 2. Overview of the approach: The camera data gets processed to be represented in the same reference frame as the GPS coordinates. The GPS coordinates from the eBike and the processed pedestrian's position is used to predict a possible crash.

information can be combined to predict possible crashes. Based on the results, the robot could interfere by either interacting with pedestrians via social interaction and/or by sending messages to approaching vehicle (e.g. a request to slow down).

In this ongoing work, we focus on the aspect of using the robot for crash prediction. We use a humanoid robot called ARI (v2) from pal robotics[1], a HNF Nicolai UD4 All-Terrain e-bike[2] that has been equipped with an V2X On-Board Unit (OBU v5.0) from Herman[3]. As shown in Figure 1, we use the following test scenario: ARI is placed at one side of a crossing and a pedestrian is crossing from the other side (coming towards ARI), while a cyclist on the eBike comes from ARI's left side towards the pedestrian. First, the robot collects visual data on the pedestrian's movement and V2X data from nearby vehicles (here the eBike). Next, these datasets are processed to extract spatial and temporal information of both agents. Finally, the system compares pedestrian's and vehicle's trajectory to identify potential crashes. This predictive model generates scenarios where trajectories either intersect (indicating potential collisions) or diverge (avoiding collisions).

[1]https://pal-robotics.com/
[2]https://hnf-bikes.com/
[3]https://www.herman.cz

II. Methods

A. Data Set

The dataset used for the collision prediction model is created through a controlled setup. The goal is to record interactions between pedestrians and vehicles in a shared space, specifically focusing on pedestrian's movement captured by the robot's cameras and vehicle data transmitted via V2X communication. ARI is placed at a known GPS location in the test environment (Figure 1). The robot's field of view captures pedestrians crossing the street while the eBike moves perpendicularly to their path. The robot records with its front RGB-D camera. Each time the pedestrian crosses, a video recording is started, capturing the pedestrian's movement. Simultaneously, the V2X data from the eBike's OBU is transmitted continuously and saved separately. This data includes Cooperative Awareness Messages (CAMs) [1], containing information about the bike's position and other relevant information. After first inspection, 8 videos capturing the full test scenario involving both the pedestrian and the eBike were retained, other videos were discarded due to different reasons like poor field of view or interruptions.

B. Pipeline

An overview of the crash prediction pipeline is shown in Figure 2. The crash prediction model requires the pedestrian's trajectory GPS_{ped} and vehicle's trajectory GPS_{veh} in the form of GPS coordinates as input. The vehicle's trajectory is already available in GPS coordinates, the RGB-D data is processed in four steps:

b) The depth model is implemented as proposed in [9]. It is based on a deep neural network and generates a depth map D_t of each video frame F_t and calculates the distance of the moving agent from the camera.

c) The pose model relies on the OpenPose tool for body and hand pose estimation [10], [11]. It extracts visual pose keypoints, joints of the human body, and provides the pixel coordinates (u, v) of all the detected keypoints k. For each pedestrian detected p_i, according keypoint coordinates are calculated.

d) The tensor combines both pixel coordinates (v, u) and depth values $d(v, u)$ of all keypoints to create a three-dimensional (3D) representation of the pedestrian's pose. For each pedestrian p_i a tensor values T_i is retrieved. In order to smoothen jumps in depth values from one frame to another a Kalman filter is applied.

e) The 3D coordinates of the keypoints are each converted into relative camera coordinates and then to real-world GPS coordinates. The average GPS coordinates are calculated to provides an accurate estimation of the pedestrian's position, GPS_{ped}.

The distance between the pedestrian and the vehicle can be computed using the haversine formula. This allows to determine which situations are at risk of a crash by checking if the calculated distance is below a certain threshold. As the dataset does not provide real crash data, a crash is simulated by forwarding the pedestrian's position GPS_{ped} at time t to a future time $t + S$, where S is chosen in a way that the trajectories intersect.

III. Results & Evaluation

Fig. 3. Visualization of calculated trajectory for one video (red: keypoints of the pedestrian and eBike, green: frame number as timestamp reference)

TABLE I
DISTANCE BETWEEN VEHICLE AND PEDESTRIAN STARTING FROM THE MOMENT THE VEHICLES ENTERS THE ROBOT'S FIELD OF VIEW. DISTANCES BETWEEN A CERTAIN THRESHOLD (HERE 5.5M) TRIGGER A CRASH WARNING.

k in t_k	Distance between vehicle and pedestrian (m)
68	5.99
73	5.98
78	5.90
83	5.93
88	6.06
93	6.29
98	6.56
103	10.20
108	6.01
113	5.90
118	5.18 (DANGER: CRASH)
123	5.32 (DANGER: CRASH)
128	5.89
133	5.74
138	5.54
143	5.38 (DANGER: CRASH)
148	5.67
153	5.43 (DANGER: CRASH)
158	5.45 (DANGER: CRASH)
163	5.45 (DANGER: CRASH)
168	5.42 (DANGER: CRASH)
173	5.33 (DANGER: CRASH)
178	5.30 (DANGER: CRASH)
183	5.27 (DANGER: CRASH)
188	5.49 (DANGER: CRASH)
193	5.76
198	6.42
203	6.53
208	6.75
213	7.75
218	7.87
223	7.98
228	9.40
233	9.92
238	10.87

The implemented approach successfully provides accurate trajectories in **5 out of 8 videos** that contain the full test scenario. Figure 3 visualizes the calculated trajectories of the vehicle and the pedestrian displayed on the frame F at time t_0. The trajectory of the eBike is already predicted, even when the bike is outside the robot's field of view. In this case, the system anticipates a potential collision at time t_{150}, where both paths appear to intersect. This demonstrates the effectiveness of a multi-modal prediction approach, where each data input contributes to reducing uncertainty in the prediction process.

However, the system still has limitations, such as inaccuracies in the GPS data. As Table I shows the distance between the pedestrian and the eBike, recorded every fifth frame, starting from the moment the vehicle enters the robot's field of view. The recorded distances range from approximately 5.99 meters to over 10 meters. For this test case, the distance threshold considered as a risk was 5.5 meters. These values appear inaccurate, especially when considering the dimensions of the shared-space where the dataset was created.

IV. Discussion

This work provides a framework that integrates the necessary components needed to combine V2X data with visual data for crash prediction. However, the current framework exhibits numerous limitations and faces several challenges.

One fundamental challenge is the accuracy of GPS data, particularly in small or confined areas. GPS performance can degrade significantly due to signal reflections, where signals bounce off nearby buildings or surfaces before reaching the receiver, introducing delays and positional errors. Blocked signals caused by obstacles such as tall buildings or dense trees can further reduce accuracy. Additionally, in smaller spaces, the receiver may have fewer satellites in view, making it harder to get precise location data. Weather conditions and the limitations of standard GPS devices, which typically have an accuracy of just a few meters, can make these issues worse. These factors make it challenging to get reliable ground truth data in small areas, especially when accurate pedestrian tracking is required. One possibility to address this challenge, is to use GPS only to issue warnings when the vehicle is at a larger distance ($\sim$20-30 m) and rely on other sensors (e.g. radar and ultrasonic sensors) in short range. Possibilities to improve the GPS accuracy in the future include RTK-GPS and sensor fusion using local SLAM.

Different shortcomings revolve about the pedestrian trajectory prediction. At this stage, pedestrian trajectory predictions were purely simulated. The current system relies on calculated ground truth data to identify potential crashes. Available models for predicting like Convolutional Transformer [11] could be used. However, these models rely on large data sets to be trained on; the small preliminary data set of 8 (usable) videos that we collected falls significantly short of the required data set size to train such a model. While possible next steps include the collection of a larger data set, it will not be feasible to collect enough data feasible for deep learning approaches. Another problem is the missing real-time capability of the system; currently, the calculation time is on the order of a few seconds. In order to be actually useful in traffic, the calculation time needs to be reduced to the order of a tenth of a second. The depth estimation sometimes lead to trajectories that showed jumps. Although this problem was significantly improved by the use of a Kalman filter, this problem might be investigated further. Finally, in the current state, the robot functions primarily as an interface between the vehicle data and pedestrian motion data streams. This functionality could be provided by stationary infrastructure too. In order to fully utilize the robot's capabilities, one next step is to enable the robot to interact directly with pedestrians. For example, the robot could provide real-time warnings or take preventive actions to avoid potential collisions. These types of interaction would be a significant advancement, moving the system closer to practical, real-world applications in shared spaces.

V. Conclusion

In this work a novel approach for multi-modal crash prediction using a social robot that relies on V2X and visual data is proposed. A framework is presented that integrates the basic functionality required to implement such an approach. The system successfully transforms pedestrians' and vehicles' trajectories in a common coordinate frame and accurately identifies potential collisions. This work demonstrates the real-world feasibility of the proposed approach. The preliminary findings indicate that such a system offers to potential to improve road safety by predicting and mitigating collision risks, contributing to safer shared spaces between pedestrians and intelligent vehicles. Shortcomings and inherent challenges provide interesting directions for further research. In particularly, the pedestrian trajectory prediction needs refinement in terms of using a causal system instead a non-causal one and real-time capabilities. Further, challenges due to GPS inaccuracy need to be addressed. Another promising direction, is advancing and exploiting the robot's interaction capabilities. While our work has the application scenario with a V2X enabled robot in mind, the presented approach is applicable beyond the setting with the robot. For instance, infrastructural solutions with camera and V2X could be installed at crossings in Cooperative Intelligent Transport Systems (C-ITS). In summary, the presented approach offers promising novel directions to increase road safety.

References

[1] A. Festag, "Cooperative intelligent transport systems standards in europe," *IEEE Communications Magazine*, vol. 52, no. 12, pp. 166–172, 2014.

[2] I. Soto, M. Calderon, O. Amador, and M. Urueña, "A survey on road safety and traffic efficiency vehicular applications based on C-V2X technologies," *Vehicular Communications*, vol. 33, p. 100428, Jan. 2022.

[3] O. A. Molina, E. Ronelöv, K. Boustedt, J. Blidkvist, and A. Vinel, "Protection of vulnerable road users using hybrid vehicular networks," in *2022 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, 2022, pp. 1–6.

[4] P. Rahimian, E. E. O'Neal, J. P. Yon, L. Franzen, Y. Jiang, J. M. Plumert, and J. K. Kearney, "Using a virtual environment to study the impact of sending traffic alerts to texting pedestrians," in *2016 IEEE Virtual Reality (VR)*, Mar. 2016, pp. 141–149.

[5] P. Rahimian, E. E. O'Neal, S. Zhou, J. M. Plumert, and J. K. Kearney, "Harnessing Vehicle-to-Pedestrian (V2P) Communication Technology: Sending Traffic Warnings to Texting Pedestrians," *Human Factors*, vol. 60, no. 6, pp. 833–843, Sep. 2018.

[6] M. Bied, B. Bruno, and A. Vinel, "Autonomous vehicles as social agents: Vehicle to pedestrian communication from v2x, ehmi and hri perspectives," in *2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2024, pp. 86–91.

[7] M. Schrapel, M. Bied, B. Bruno, and A. Vinel, "Experiencing social robots for traffic guidance using virtual reality videos," in *Mensch und Computer 2024 - Workshopband*. Gesellschaft für Informatik e.V., 2024.

[8] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *International Journal of Human-Computer Studies*, vol. 77, pp. 23–37, May 2015.

[9] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman, "Learning the depths of moving people by watching frozen people," 2019. [Online]. Available: https://arxiv.org/abs/1904.11111

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[11] K. Chen, H. Zhu, D. Tang, and K. Zheng, "Future pedestrian location prediction in first-person videos for autonomous vehicles and social robots," *Image and Vision Computing*, vol. 134, p. 104671, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885623000458