Evaluation of SIMMARC: An Audiovisual System for the Detection of Near-Miss Accidents

Florian Krebs[1(✉)], Georg Thallinger[1], Helmut Neuschmied[1], Franz Graf[1], Georg Huber[2], Kurt Fallast[2], Peter Vertal[3], and Eduard Kolla[3]

[1] Joanneum Research, Graz, Austria
florian.krebs@joanneum.at
[2] Planum, Graz, Austria
[3] University of Žilina, Žilina, Slovakia

Abstract. In this paper, we present and evaluate a system that automatically identifies hazardous traffic situations using visual and acoustic sensors. The system has been installed at three locations in Austria and several months of audio and video data have been analyzed. We evaluate the accuracy of the employed data analysis algorithms as well as the usefulness of the detected events for the overall task of assessing the risk potential of a road intersection. Our results show that the long-term analysis made possible by the proposed system leads to a better understanding of the risk potential of traffic areas, and can finally serve as a basis for defining and prioritizing improvements.

Keywords: Near-miss accidents · Accident detection · Automatic event detection

1 Introduction

Nearly 1.35 million people die in road accidents each year, according to the annual report of the World Health Organization [1]. The reduction of fatalities is therefore one of the most important aims of humanity which drives the development of safer vehicles and road infrastructure.

Authorities in many countries have a predefined procedure for assessing and optimizing the safety of their road infrastructure. For example, the Austrian research association for roads, railways and transport (FSV[1]), defines a crossing or street section as *accident black spot*, if either at least three similar accidents with bodily injury have occurred within three years, or at least five accidents with bodily or material damage have occurred within one year[2]. Once such an accident black spot is identified, the FSV suggests carrying out a safety inspection of the involved road section to investigate the underlying causes of the accidents.

[1] http://www.fsv.at/.
[2] RVS 02.02.21.

A. L. Martins et al. (Eds.): INTSYS 2019, LNICST 310, pp. 195–202, 2020.
https://doi.org/10.1007/978-3-030-38822-5_13

As this is a purely reactive approach, where accidents must occur before a black spot is identified, methods were proposed to identify dangerous traffic spots based on near-miss scenarios [2, 3, 8]. These systems are able to detect near-miss scenarios semi-automatically by analyzing data gathered at a traffic area. Suspicious events are identified automatically and then presented to a human expert to decide whether the detected events are relevant and which actions to take.

In this paper, we present an evaluation of the audiovisual analysis system proposed in [2], after analyzing three traffic spots in Austria over a period of several months. Once the methodology has been reviewed, we outline the strengths and limitations of the system and present ideas for future improvements.

2 System Overview

In this section, we shortly describe SIMMARC [2], a system for the audiovisual detection of dangerous scenes. The system gathers audio and video traffic data from sensors that are installed on poles at the traffic area (see Fig. 1).

Fig. 1. Camera and microphone installation at Wickenburggasse (left) and Dietrichsteinplatz (right) in Graz, Austria.

The sensor data is then used to detect the following events in real-time:

- Emergency braking actions (video)
- Car horns (audio)
- Tire squealing (audio)
- Tram bells (audio)

Once an event is detected, the corresponding audiovisual footage (including 30 s before and after the event) is captured to an incident store. This incident store is periodically analyzed by a traffic expert.

In the following, we describe the methodology to extract the events mentioned above.

2.1 Detection of Audio Events

For the detection of acoustic events we have selected a recurrent neural network due to its computational efficiency. To extract the three acoustic events (car horn, tire squealing, tram bell) from the raw audio signal, the signal is segmented into overlapping time frames (length 46 ms) to obtain a frame rate of 50 fps. From each time frame, we compute the logarithm of the mel-filtered magnitude spectrogram. We chose a filter bank with 120 mel frequency bands between 200 and 16000 Hz, in order to capture all important harmonics of the target sounds. These features are fed into a recurrent neural network, which consists of two Gated Recurrent Unit [9] layers with 30 hidden units each and a final classification layer which outputs a scalar that indicates the presence of the corresponding audio event. We apply exponential smoothing to the activation function with a time constant of 0.24 s. Once the smoothed activation function exceeds a certain threshold, an event is considered to be present.

2.2 Detection of Video Events

The aim of the visual analysis is to detect the position, type (car, bus, pedestrian, etc.), speed and acceleration resp. deceleration of various road users. We divide this task into two steps: First, we recognize the rough vehicle position (bounding box) and the type of road user with the Yolo v3 neural network [7]. Then, we determine speed, acceleration and a refined position (the used traffic area) with the feature-based point tracking method proposed in [5]. Distinguishable points in the image are tracked and an algorithm similar to [6] is used to cluster the resulting trajectories and assign them to individual objects. The outcome of the whole process is illustrated in Fig. 2 and the process is described in the following.

To determine the object velocity, the first step is to calculate the velocity of the object points with the assumption that they are all at ground level. In reality, however, points have different heights which yields different velocities for the same object. These speed differences are used to calculate the actual height of the points and thus correct their velocities. With plausibility checks (e.g. the height of the object must be above the road level) and by rejecting outlier points, the point velocities are refined and merged into an object velocity. As an additional result we get a 3D point cloud of each object, from which the used traffic area (the projection of the vehicle onto the ground plane) can be calculated. In order to ignore shadows, points near the ground plane are discarded (see Fig. 2).

The speed of an object is then used to calculate the acceleration of a traffic user. If the braking acceleration falls below -4 m/s^2, the corresponding time is reported as emergency braking.

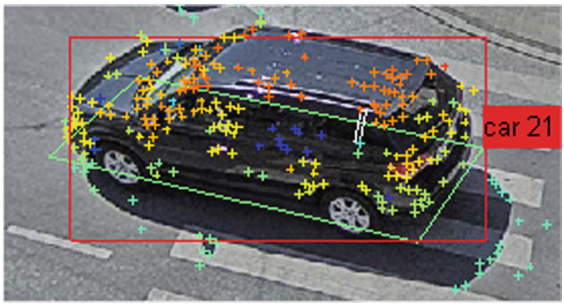

Fig. 2. In video images, objects are detected by the Yolo v3 detector (red rectangle with the object label). From the tracked feature points the speed, acceleration, and the used traffic area (green rectangle) are calculated. (Color figure online)

3 Evaluation

3.1 Data

In order to evaluate the system, data was recorded at three locations in Austria (see Fig. 3):

Graz, Dietrichsteinplatz. The Dietrichsteinplatz node (Fig. 3 left) is a heavily frequented, unregulated intersection, with peak loads of 1,000 vehicles/h at the evening on working days. The diversity of traffic participants (tram, regional bus, taxi, car, truck, bicycle, foot), the cramped space conditions and the short distance to a neighboring light signal-controlled intersection have repeatedly led to structural and traffic-related adaptations in recent decades. In total, we recorded 5 days in June 2017.

Graz, Wickenburggasse/Körösistraße. The highly frequented, signal controlled junction (Fig. 3 center) features an intersection of two main cycle routes. The sum of the access loads in the morning peak on working days is around 1,800 vehicles/h and around 900 cyclists/h. In total we recorded 52 days between April and July 2018.

Velden, Kärntner Straße. This location (Fig. 3 right) is a shared space. A large number of pedestrians use this area, especially during the summer months. We recorded 15 days in August 2018.

Fig. 3. Camera view of the recording locations (from left to right: Dietrichsteinplatz, Wickenburggasse, Kärntner Straße).

At each location, one microphone and one camera were installed on a pole at approximately 6 m and 10 m height respectively. In total, we recorded 72 days of video and audio footage.

3.2 Detection of Audio Events

In the following the audio detection algorithms are evaluated.

Datasets. We randomly sampled 32 h of the recorded data for training, 8 h for validating, and 34 h for testing the algorithms and annotated the occurrence of the three target audio events (horn, tire squealing, and tram bell). We made sure that there is no overlap between the three sets.

Evaluation Metrics. In contrast to other sound event detection evaluations we are not interested in the exact start and stop time of an event. For our application it suffices to know whether an event has occurred within a certain temporal window. Therefore, we use the following approach: An event is counted as true positive if there is at least one frame overlap between a detected and an annotated event. If a detection occurs outside of an annotated event, it is counted as false positive. An event is counted as false negative, if all detections of the corresponding event are below the threshold. Then, we compute three metrics: F-measure, Precision, and Recall [10]. The threshold was determined by maximizing the F-measure on the validation set.

Table 1. Audio event classification results on the 34 h test set.

Event	F-measure	Precision	Recall
Horn	0.76	0.89	0.66
Tire squealing	0.51	0.47	0.56
Tram bell	0.81	0.85	0.83

Results. The results on the test set are shown in Table 1. In total, there were 61 horn events, 21 tire squealing events, and 17 tram bell events in the 34 h test set. As can be seen, the tram bell detector was found to perform best, probably because the tram bell sound does not vary much and the sound stands out from the remaining sounds. The most common (false positive) errors of the detectors are:

Horn detector: Ambulance siren, shouting children, brake squealing, tram bell
Tire squealing: Car brake squealing, children squealing
Tram bell: Children squealing

3.3 Detection of Braking Actions

Groundtruth. In order to assess the accuracy of the visual speed and acceleration measurements, we selected 11 scenes which cover a variety of dangerous situations. A 3D model of the scenes including the vehicles was constructed by mapping

simulated vehicles onto the video recordings using the software PC-Crash[3]. Once a 3D model of a scene was constructed, the speed and acceleration trajectories of all traffic participants was obtained and compared to the output of the video analysis system. Measuring location and speed trajectories by 3D reconstruction is commonly used for accident reconstruction [4].

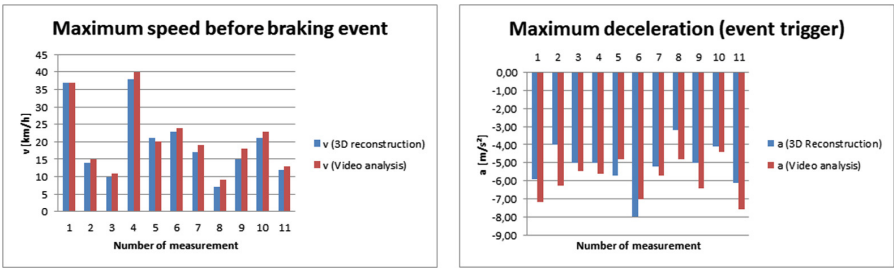

Fig. 4. Comparison of speed and deceleration values from the video analysis system and the 3D model reconstruction.

Results. Figure 4 shows the maximum speed and deceleration values before a braking event of the video analysis system and the 3D reconstructions. As can be seen, the speed measurements closely match the 3D reconstructions (mean error 1.45 km/h). However, the measurement of the deceleration is less precise (mean error 1.07 m/s^2). Since the deceleration is calculated from successive velocity values, changes in velocity measurement errors have a large influence on the error of the deceleration values. Results could probably be improved by averaging the deceleration over a longer period of time, with the disadvantage that shorter braking maneuvers would not be recognized.

3.4 Detection of Relevant Situations

In this section we evaluate the relevance of the detected scenes for improving the safety of a road section. To that means, we performed two experiments:

Experiment 1. In the first experiment, we addressed the question "*What percentage of the detected scenes is actually relevant for traffic experts?*". In this regard, the scenes detected at Dietrichsteinplatz were presented to a traffic expert, who classified them as relevant or not.

Results. In the 78 h of recording at Graz, Dietrichsteinplatz, the system detected 505 events. Of these 505 events, 430 events were detected correctly, and 39 events were identified as relevant by the traffic expert. The relevant scenes contained various disregards of traffic regulations (from drivers, pedestrians, cyclists and motorists) like ignoring the right of way or road markings, and frequent strong decelerations at parts of

[3] http://www.dsd.at.

the junction. The traffic experts concluded that the traffic participants often are not aware of the priority situation at the crossing, which could be improved by further road markings or traffic signs.

If the manual inspection of one recorded scene is assumed to take 10 s, it takes roughly 1.4 h to watch the 505 detected scenes for the observation time period of 78 h. This means a reduction to 1.8% of the original material. These numbers are expected to vary from location to location according to different traffic volumes and intersection layouts.

Experiment 2. This experiment addressed the question "*Are the detected scenes representative of the actual traffic situation?*". Therefore, a manual observation at the location Wickenburggasse/Körösistraße was carried out at the same time the automatic system was present. A representative period on a working day from 07:00 to 10:00 o'clock was selected and both video and audio was recorded. Then we compared the detections of (i) an expert on site, (ii) another expert watching the footage in the office, and (iii) the automatic detection system.

Results. The evaluation showed that the expert on site recorded 55 noteworthy incidents within 3 h, while the expert in the office identified 98 incidents. Among these, 31 incidents were detected of both experts. As the expert in the office had a restricted field of view, it can be assumed that the real agreement is higher. During the same reference period, the automatic system detected 19 incidents, 16 that were also identified by the experts, and three additional ones. Obviously, the current automatic system cannot detect incidents that do not coincide with braking or acoustic cues (e.g. bikes crossing the street at red traffic light). Therefore, we plan to extend the set of detected events in future work. Nevertheless, the automatic system outperforms the human eye in detecting rapid decelerations, as these happen at very short time scales.

4 Conclusions and Future Work

In this paper, we evaluated various aspects of an automatic system that detects potentially dangerous events on three road intersections in Austria. We showed that the detection of four event types (car horn, tire squealing, tram bell, and emergency braking) already yields scenes that are relevant for traffic planners to assess the risk potential of an intersection. Using a semi-automatic system, the time needed to inspect a location can be drastically reduced (in the described case to 1.8% of the original time period) and therefore enables long-term analyses. Future work will be extending the set of reported events by exploiting the position and distances between traffic participants and to automatically derive an assessment of the severity of a detected traffic event.

Acknowledgements. This research was partially funded by the Austrian Research Promotion Agency (FFG) within the program "Mobilität der Zukunft".

References

1. Global status report on road safety 2018. World Health Organization (2018)
2. Thallinger, G., et al.: Near-Miss Accidents – Classification and Automatic Detection. In: Kováčiková, T., Buzna, Ľ., Pourhashem, G., Lugano, G., Cornet, Y., Lugano, N. (eds.) INTSYS 2017. LNICST, vol. 222, pp. 144–152. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93710-6_16
3. Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M.: Audio surveillance of roads: a system for detecting anomalous sounds. IEEE Trans. Intell. Transp. Syst. **17**(1), 279–288 (2016)
4. Edelman, G., Bijhold, J.: Tracking people and cars using 3D modeling and CCTV. Forensic Sci. Int. **202**, 26–35 (2010)
5. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings DARPA Image Understanding Workshop, pp. 121–130 (1981)
6. Saunier, N., Sayed, T.: A feature-based tracking algorithm for vehicles in intersections. In: Proceedings of the 3rd Canadian Conference on Computer and Robot Vision, pp. 59–59. IEEE (2006)
7. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
8. Green, E.R., Agent, K.R., Pigman, J.G.: Evaluation of auto incident recording system (AIRS) (2005)
9. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
10. van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworth, London (1979)