Detecting Road User Actions in Traffic Intersections Using RGB and Thermal Video

Bahnsen, Chris; Moeslund, Thomas B.

Published in:
2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)

DOI (link to publication from Publisher):
10.1109/AVSS.2015.7301733

Publication date:
2015

Document Version
Accepted author manuscript, peer reviewed version

Detecting Road User Actions in Traffic Intersections Using RGB and Thermal Video

Chris Bahnsen, Thomas B. Moeslund
Visual Analysis of People Laboratory, Aalborg University
Rendsburggade 14, Aalborg, Denmark
{cb,tbm}@create.aau.dk

Abstract

This paper investigates the development of a watch-dog system that detects a subset of road user actions in traffic intersections. Footage of the intersections is captured with RGB and thermal cameras to ensure that the road is visible round-the-clock even in difficult weather conditions. The watch-dog system consists of several, cascaded detectors which are capable of detecting specific road user actions, such as Right Turning Vehicles, Left Turning Vehicles, and Straight Going Cyclists. Experimental results on 4 hours of video from 3 different intersections show good performance and a precision above 0.93 when detecting turning vehicles. The use of both RGB and thermal video generally results in better performance, providing overall stability when observing the road.

1. Introduction

It is the goal of the European Commission to cut the number of road deaths by 50 % in 2020 and diminish the number almost entirely by 2050 [3]. In order to reach these goals, not only the security of the vehicles must be enhanced but also the layout of the roads must change to enhance safety. Historically, road layouts have been changed to the better based on previous knowledge of road fatalities and deaths. This means that traffic researches and designers must wait for accidents to happen in order to improve the layout of the road.

In surrogate safety analysis, however, it is sufficient to measure the number of accidents that almost happened. The foundation behind surrogate analysis is the existence of a continuous relationship between the levels of severity of an accident and their corresponding frequencies. For instance, it is assumed that slight injuries occur more frequently than severe injuries and thus one may form a safety pyramid [6] where the fatal injuries resist at the very upper parts of the pyramid (severe, low frequency) and normal traffic fill up the bottom parts of the pyramid (normal, high frequency). By counting the number of near-accidents where a critical interaction between road users nearly happened, one achieves a *surrogate measure* of the number of more severe, fatal interactions [14]. Recently, this rationale has been taken even further by indicating that less-severe, normal traffic interactions enables traffic researches to monitor the safety level [11], [15]. This enables a rapid safety analysis of roads from data over weeks instead of years.

Special attention is needed in improving the safety of vulnerable road users (VRU). VRUs is defined as pedestrians, elderly, disabled persons, cyclists, and riders of powered two-wheelers (mopeds and motorcycles). Compared to the total number of traffic accidents, VRUs account for a disproportionately high number of road fatalities and injuries. In 2013, according to the European Commission [3] more than 14.000 VRUs were killed in the European Union. It is the long-time goal of this project to enable traffic researches to improve the safety of VRUs by gaining knowledge of the accident causations. In this work, we are laying the foundation by studying specific movements of selected road users at intersections.

1.1. Monitoring road users

We have to study the roads in order to understand the frequency and nature of accidents and near-accidents. Manually monitoring the roads is tedious and inflexible and does not allow for a larger understanding of accident causation. A more flexible approach is to record the roads with a camera and watch the footage off-line. This allows for the reconstruction of critical events but still presents the user with a tremendous amount of data. The optimal solution to this problem is to design a system that automatically detects and tracks the road users from the recorded video data. From these tracks, traffic analysts can define heuristics that determine the interactions between the road users and on a higher level, the safety of a particular road.

However, the detection and tracking of objects in uncon-

strained scenes is still an unsolved problem. State-of-the art tracking systems are usually evaluated at 2-minute intervals under static weather conditions and does not perform well under occlusion, clutter, and illumination changes. No tracker is currently capable of detecting and tracking objects round-the-clock in unconstrained scenarios. Smeulders et al [13] provides a good review and performance evaluation of recent trackers.

Jackson et al [7] have developed an open-source toolbox for traffic video analysis which forms the base line for surrogate traffic analysis such as the work of [12]. However, as with general tracking algorithms, the length of the dataset is short and the video data is captured under good weather conditions [10]. The toolbox builds upon the popular KLT-tracker [1] which needs an additional grouping of tracked features to convert a number of tracked points to a number of tracked objects. The grouping is often ambiguous - is it a single bicycle or a cluster of cyclists waiting at the stop line? Is it a truck with a trailer - or two separate vehicles?

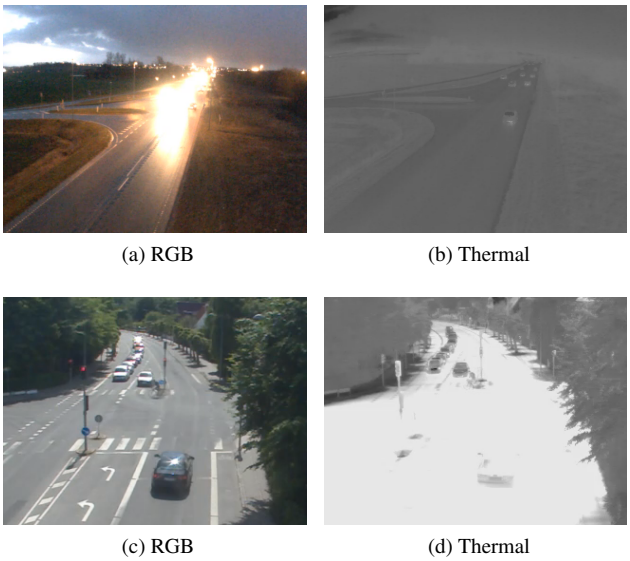
1.2. Reducing the amount of video

Because tracking still remains an unsolved problem, we acknowledge that there is a need for a human-in-the-loop to assess the nature and severity of the events between road users. However, we may design a system that reduces the amount of video data to manually assess. Such a *watch dog* should not necessary track all road users at all times but instead detect whenever there are situations that need further investigation - and whether there are periods of time when nothing of interest occurs. In this work, we will refrain from detecting interactions between road users but instead study the individual actions of the road users and obtain a reliable detection. Once these detections are achieved, one may obtain interactions by combining the detections. We build upon the ideas introduced by Madsen et al [9]. In this work, we introduce a thorough evaluation of the individual detectors on novel datasets in both the RGB and thermal domains. Furthermore, we explain the algorithmic framework behind the detectors and how they are enhanced to work in both domains.

The issue of observing the road through a camera is treated in Section 2. The proposed watch dog that operates on the video data is presented in Section 3. Experiments are discussed in Section 4 and concluding remarks are presented in Section 5.

2. Observing the road

Even the best tracking systems are only as good as the data they process. We want to detect the road users round-the-clock in all weather conditions which means that the road should be observable in almost any condition by looking at the recordings provided. Traditional surveying techniques employ one or multiple visible light (RGB) cameras to monitor the intersection [8]. While this works well under good weather conditions, the video data still stuffer from varying shadows and very sparse information during the night. Thermal cameras, on the other hand, capture the radiated heat from objects and are thus not sensitive to changes in the environment as long as the object of interest has a different temperature from the background. For a survey of thermal cameras, refer to [5]. However, the thermal modality is poor on features which makes it much harder to discriminate between objects, recover identities after occlusion, or classify road users. Together though, RGB and thermal cameras supplement each other and extend the visibility of the road. In this work, we use a joint configuration of a RGB and thermal camera to monitor road intersections. See Figure 1 for a comparison of the two modalities.



(a) RGB (b) Thermal

(c) RGB (d) Thermal

Figure 1: (a), (b): RGB and thermal images of an intersection at dusk. In the RGB modality, the headlights of the cars passing by dominates most of the road. In the thermal modality, the cars are fully visible. (c), (d): RGB and thermal images of an intersection in full sunlight. In the thermal image, the car on the right is barely visible due to the heated asphalt whereas the biker on the upper pedestrian crossing stands out. The RGB image is fully visible.

3. Watch-dog system

Because our system should be able to function as a watch-dog to a human operator, robustness to changes in the environment is more important than the ability to perfectly detect and track road users. In order to make the watch-dog robust, we tailor the system to perform a number of specific tasks in certain areas of interest. We use the geometry of the

intersection to infer specific patterns that road users must take to complete an action. For instance, if we want to analyse a vehicle doing a right-turn at an intersection we know that the vehicle must (1) enter the intersection, (2) perform a right turn, and (3) exit the intersection. These tasks may be solved in succession:

1. **Detect presence:** Detect if an object is present at the chosen entry point of the intersection. If the size of the object fulfils the criteria of the vehicle type, proceed to step 2. Otherwise, discard the object.
2. **Detect movement:** Detect if the object of interest is turning right, e.g. if there is movement in a certain direction in a predefined area of the intersection. If the movement is sufficient, proceed to step 3.
3. **Detect presence:** Detect if the object is present at the chosen exit point of the intersection by applying the method of step 1.

We assume that a vehicle has made a right turn if the three tasks are completed in succession. If not, the vehicle is doing something else - which another detector may detect.

In this specific context, we create the foundation to detect near-conflicts between vehicles and cyclists at urban signalized traffic intersections. In order to do so, we want to detect right turning vehicles, left turning vehicles, and straight going cyclists. The three detectors all consists of a chained combination of the two basic tasks; *detecting presence* and *detecting movement* which are further described in the following.

3.1. Detecting presence

When detecting presence, we want to detect if a road user is present or not at a given region of interest in the image. This is obtained via a background subtraction technique applied to the specific region of interest (ROI). We use a background subtraction technique based on reference images which are updated according to the routine described below:

1. Perform Canny edge detection [2] on current image and obtain edge image.
2. Subtract edge image from background edge image.
3. Filter noise.
4. Find pixel sum of remaining edges. If sum is above threshold, the detector is triggered.
5. Update background if the following criteria are satisfied:
   (a) Motion between current and previous frame is below 10 % of threshold for τ_1 concurrent frames.
   (b) Pixel sum is below 80 % of threshold, and background has not been updated for τ_2 consecutive frames.

The routine above is applied independently on both the RGB and thermal modality. The threshold is found experimentally for each intersection and modality and is higher when detecting vehicles than detecting bicycles due to the difference in size of these road users.

3.2. Detecting movement

Estimation of the movement in a ROI of the video is obtained by using the two-frame dense motion estimation of Farneback [4] with the following procedure:

1. Calculate the dense optical flow of the ROI.
2. Count number of flow vectors of certain magnitude inside a chosen flow range.
3. Threshold vector count and update confidence measure.

The flow range mentioned in step 2 is chosen to only detect movement in the preferred range of the detector. For instance, we only want do detect movement from left to right when detecting right turning vehicles.

3.3. Chaining actions

It is of special interest of the traffic researchers to know whenever a road user is stationary in certain areas of the intersection. Therefore, we combine the tasks of detecting presence and movement into a third detector, the stationary object detector. The stationary object detector is triggered whenever something is present within the ROI and there is no or little movement, or flow, inside the ROI.

As described at the beginning of Section 3 we define events inside the intersection by chaining sequential actions. By tailoring the detectors for specific needs we focus the overall generic tracking problem to solve a very constrained problem at hand. Other problems, for instance right turning cyclists, might be solved by building another chained set of detectors. The task of detecting right turning vehicles is performed by the use of five detectors; two presence detectors, abbreviated E, two movement detectors (F), and one stationary detector (S). The number of detectors used for detecting left and right turning vehicles, and straight going cyclists is listed in Table 1.

A vehicle is detected as a right turning candidate whenever it enters the entry point of the intersection which is laid out in the ROI of detector E1 (Figure 2a). Whenever detector E1 is triggered, the movement detector F1 and the stationary detector S1 are activated. The detector F1 looks for movement in the direction of the arrow (see Fig. 2a) and

	RTV	LTV	SGC
Detecting presence (E)	2	2	3
Detecting movement (F)	2	4	1
Stationary object (S)	1	0	0

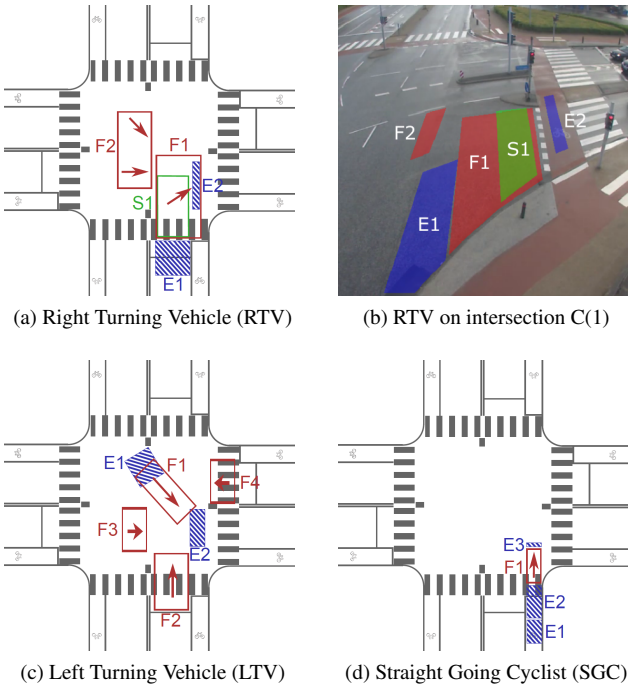
Table 1: Detector types, and their shorthand notation, used when detecting Right Turning Vehicles (RTV), Left Turning Vehicles (LTV), and Straight Going Cyclists (SGC).

detector S1 detects if the vehicle has stopped. If F1 has detected that the vehicle is turning, the detector E2 is activated to judge if the vehicle enters the conflict zone which concludes the detection. If S1 is activated, we assume that the vehicle has stopped in the middle of the intersection and is possibly awaiting clearance to turn. In this case, we let the other detectors stay open a little longer to detect an eventual turn of the vehicle. If no action occur in the detectors E2, F2, and S1, they are deactivated after a short duration of time. The detector F2 is used to filter out false positives, for instance vehicles going from left to right in the intersection. An activity diagram explaining the work-flow of the Right Turning Vehicle (RTV) detector is shown in Figure 3. The RTV detector is shown on an intersection prototype in Figure 2a and in an actual configuration in Figure 2b.

The Left Turning Vehicle (LTV) detectors and Straight Going Cyclist (SGC) detectors work similarly to the RTV detector. In the LTV detector, the stationary detector is discarded and the area of the presence detector (E1) is moved further into the intersection. Two movement detectors (F3, F4) have been added to filter out false detections from vehicles from other directions, complementing the F2 of the RTV detector. The proposed layout of the LTV detector is shown in Figure 2c. The SGC detector adds one presence detector to help filter pedestrians and cars from cyclists. It discards the detectors F2, F3, and F4 as they have shown to be of little use in this specific case. The SGC detector prototype is seen in Figure 2d. A straight going cyclist is detected if the detector E3 is activated in a chain of actions.

3.4. Fusing modalities

The video data of the intersections is captured by both a conventional RGB and a thermal camera. In this experiment, we synchronize the two modalities and run the detectors on each modality concurrently. Each underlying detector, i.e. the presence and movement detector, operates on both a RGB and a thermal image. For each modality, the detector outputs a confidence value between 0 and 1. An individual detector is triggered if the confidence is above 0.5. A multi-modal detector must have an averaged confidence value above 0.5 to be triggered.



(a) Right Turning Vehicle (RTV)

(b) RTV on intersection C(1)

(c) Left Turning Vehicle (LTV)

(d) Straight Going Cyclist (SGC)

Figure 2: RTV, LTV, and SGC detectors on intersection prototypes

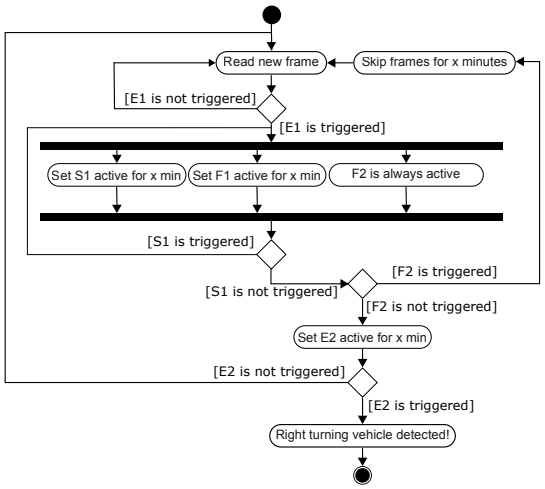


Figure 3: Activity diagram of the Right Turning Vehicle (RTV) detector.

4. Experimental results

The RTV, LTV, and SGC detectors are evaluated at three different intersections located in the Danish cities of Aalborg (A, B) and Viborg (C). The duration of the evaluated video data is four hours in total. The data is captured in the morning peak hour to capture as much traffic as possible and thus challenge the algorithms. The conditions of the

evaluated intersections are listed in Table 2. Samples from the intersections are shown in Figure 4.

Intersection	Time	Weather	Temperature
A(1)	07:00 - 08:00	Sunny	13 °C
A(2)	07:00 - 08:00	Overcast	15 °C
B(1)	07:00 - 08:00	Rain	12 °C
C(1)	07:00 - 08:00	Overcast	13 °C

Table 2: Conditions of the evaluated video data. Video A(2) is showing the same intersection as A(1), four days later.

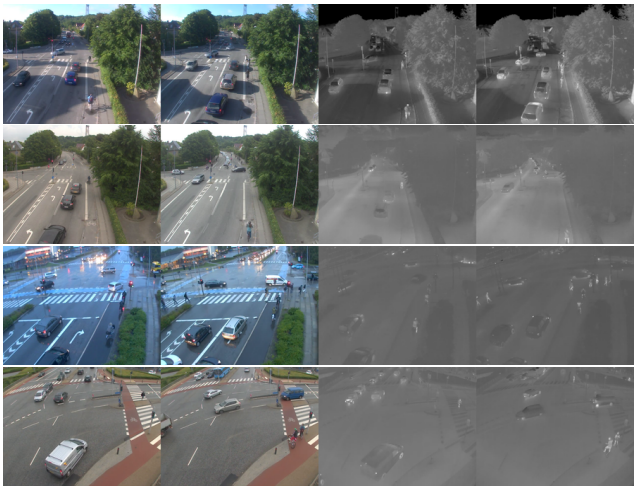


Figure 4: Snapshots of the intersections used in the experiments. For each intersection, two frames are shown in both the RGB and thermal modalities. From top to bottom; A(1), A(2), B(1), C(1).

For each of the locations, right turning vehicles, left turning vehicles, and straight going cyclists have been annotated manually and assigned a time stamp which corresponds to the entry of the vehicle or cyclist in the final presence detector (E2/E3) of the RTV, LTV, and SGC detectors. The detectors are fitted to each of the intersections using the first ten minutes of video. For sequence A(1) and A(2), the same settings are used. A detection is considered a true positive if its time stamp is within ± 2 seconds of the nearest ground truth time stamp. Detections and the corresponding ground truths can only be associated once, i.e. only one of multiple detections may be marked as a true positive if they all correspond to the same ground truth label. The results of the experimental evaluation are listed in Table 3. The detectors are evaluated on the RGB and thermal modalities both separately and combined.

Overall, the results show good performance of the RTV and LTV detectors, resulting in a precision of 0.94–1.00 and a recall of 0.80–0.97 when combining both modalities (RGBT). The SGC detector performs poorer than the RTV and SGC in the four sequences, most notably in the RGB modality. The poorer performance of the cyclist detection is possibly due to occlusion and the case that cyclists riding side-by-side are detected as a single cyclist. Cyclists are more distinguished in the thermal modality which is reflected by higher precision rates than the corresponding RGB detections.

In 15 out of 24 cases (precision+recall), the detectors operating on RGBT perform better than or equal to the best performing single modality. In the remaining 9, the performance is better in a single modality. However, in these cases, the RGBT is trailing behind the best performing modality by typically 0.01–0.03, even if the other single modality performs considerably worse.

5. Conclusions

This work presented a system that detects right and left turning vehicles, and straight going cyclists in signalized intersections by using RGB and thermal video data. It does so by chaining the output of two fundamental detectors which detects presence and movement. The spatial constraints of the intersections are used to create chains of actions that classifies a road user. The detectors are evaluated on a total of four hours of data from three different intersections. The results are promising and shows that the combination of RGB and thermal video may lead to a more stable detection of the road users in real-life, long-term traffic video.

Future work includes a more sophisticated fusion of the modalities by using contextual information to create a confidence measure reflecting the reliability of a modality. Furthermore, the detections will be combined to produce an estimate of the interactions between road users at the selected intersections.

Acknowledgements

The authors thank Tanja Kidmann Osmann Madsen for acquiring the data as well as providing the ground truth. This research was supported by a grant from the European Commission under the Horizon 2020-programme, H2020-EU.3.4.

References

[1] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004. 2

[2] J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):679–698, 1986. 3

[3] European Commission. White paper roadmap to a single european transport area towards a competitive and resource efficient transport system. *COM (2011)*, 144, 2011. 1

Intersection A(1) (1 hour)

	TP			FP			FN			Precision			Recall		
	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV
RGB	**103**	61	65	420	57	5	**43**	61	12	0.20	0.52	0.93	**0.71**	0.50	0.84
T	102	92	50	31	**2**	**3**	44	30	27	0.77	**0.98**	0.94	0.70	0.75	0.65
RGBT	**103**	**97**	**71**	**25**	3	**3**	**43**	**25**	**6**	**0.80**	0.97	**0.96**	**0.71**	**0.80**	**0.92**

Number of positives: SGC 146, RTV 122, LTV 77

Intersection A(2) (1 hour)

	TP			FP			FN			Precision			Recall		
	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV
RGB	92	**108**	**90**	43	2	9	49	**12**	**3**	0.68	0.98	0.91	0.65	**0.90**	**0.97**
T	91	102	15	16	2	**1**	50	18	78	0.85	0.98	0.94	0.65	0.85	0.16
RGBT	**97**	**108**	83	**11**	**0**	3	**44**	**12**	10	**0.90**	**1.00**	**0.97**	**0.69**	**0.90**	0.89

Number of positives: SGC 141, RTV 120, LTV 93

Intersection B(1) (1 hour)

	TP			FP			FN			Precision			Recall		
	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV
RGB	**52**	237	175	99	**8**	**1**	**19**	116	35	0.34	**0.97**	**0.99**	**0.73**	0.67	0.83
T	48	276	144	26	51	3	23	77	66	0.65	0.84	0.98	0.68	0.78	0.69
RGBT	48	**301**	**200**	**24**	16	5	23	**52**	**10**	**0.67**	0.95	0.98	0.68	**0.85**	**0.95**

Number of positives: SGC 71, RTV 353, LTV 210

Intersection C(1) (1 hour)

	TP			FP			FN			Precision			Recall		
	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV	SGC	RTV	LTV
RGB	**54**	**109**	**94**	6	**7**	**3**	**20**	**7**	**5**	0.90	**0.94**	**0.97**	**0.73**	**0.94**	**0.95**
T	49	106	80	6	14	**3**	25	10	19	0.89	0.88	0.96	0.66	0.91	0.81
RGBT	52	108	93	**2**	**7**	4	22	8	6	**0.96**	**0.94**	0.96	0.70	0.93	0.94

Number of positives: SGC 74, RTV 116, LTV 99

Table 3: Detection performance of the RTV, LTV, and SGC detectors evaluated at four different video sequences. A detection is marked as a true positive if it is within $\pm$ 2 seconds of the ground truth.

[4] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, pages 363–370. Springer, 2003. 3

[5] R. Gade and T. B. Moeslund. Thermal cameras and applications: A survey. *Machine vision and applications*, 25(1):245–262, 2014. 2

[6] C. Hydén. The development of a method for traffic safety evaluation: The swedish traffic conflicts technique. *BULLETIN LUND INSTITUTE OF TECHNOLOGY, DEPARTMENT*, (70), 1987. 1

[7] S. Jackson, L. F. Miranda-Moreno, P. St-Aubin, and N. Saunier. Flexible, mobile video camera system and open source video analysis software for road safety and behavioral analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2365(1):90–98, 2013. 2

[8] A. Laureshyn. *Application of Automated Video Analysis to Road User Behaviour*. Lund University, 2010. 2

[9] T. K. O. Madsen, C. Bahnsen, H. Lahrmann, and T. B. Moeslund. Automatic detection of conflicts at signalized intersections. In *Transportation Research Board 93rd Annual Meeting*. 2

[10] N. Saunier, H. Ardö, J.-P. Jodoin, A. Laureshyn, M. Nilsson, Å. Svensson, L. Miranda-Moreno, G.-A. Bilodeau, and K. Åström. A public video dataset for road transportation applications. 2013. 2

[11] N. Saunier and T. Sayed. Automated analysis of road safety with video data. *Transportation Research Record: Journal of the Transportation Research Board*, 2019(1):57–64, 2007. 1

[12] N. Saunier, T. Sayed, and K. Ismail. Large-scale automated analysis of vehicle interactions and collisions. *Transportation Research Record: Journal of the Transportation Research Board*, 2147(1):42–50, 2010. 2

[13] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1442–1468, 2014. 2

[14] Å. Svensson and C. Hydén. Estimating the severity of safety related behaviour. *Accident Analysis & Prevention*, 38(2):379–385, 2006. 1

[15] A. P. Tarko. Use of crash surrogates and exceedance statistics to estimate road safety. *Accident Analysis & Prevention*, 45:230–240, 2012. 1