In-vehicle Driver and Passenger Activity Recognition

Martin Torstensson, Thanh Hai Bui, David Lindström, Cristofer Englund, and Boris Duran*

RISE Research Institutes of Sweden

May 10, 2019

Abstract

Recognition of human behaviour within vehicles are becoming increasingly important. Paradoxically, the more control the car has (i.e. in terms of support systems), the more we need to know about the person behind the wheel [1] especially if he or she is expected to take over control from automation. A lot of focus has been devoted to research on the sensors monitoring the outside surroundings, but sensors on the inside has not received nearly as much attention. In terms of monitoring distractions, what is currently seen as dangerous (e.g. use of mobile phones) can in the future be seen as something good that helps to keep people awake in highly automated vehicles. Another reason for mapping activities inside the car is the often occurring mismatch between driver expectations and the reality of what today's automated vehicles are capable of [2]. As long as the automation comes with limitations that impose a need for the driver to take over control at some point, it will be important to know more about what happens inside the vehicle. In this paper we describe the work performed within the ongoing DRAMA project[1] to combine UX research with computer vision and machine learning to gather knowledge about what activities in a cabin can be mapped how they can be modelled to improve traffic safety and UX functionality.

1 SCENARIOS INDUCED SYSTEM ARCHITECTURE

Successful mapping of driver and passenger activity has far-reaching implications for both UX and safety functions in autonomous vehicles (AVs). With correct classifications of what the driver is doing, the human-machine interaction can be directed to the most suitable modality (visual/audio/haptic) at each moment. If the car knows the full body position (sitting, lying, etc) of its passengers, the safety functions can be adapted to the in-the-moment best deployment of for example airbags, steering, brake and crash avoidance patterns. In-vehicle activity and driver attention/disengagement in the driving task can be used by the AV to decide if a hand-over can be done safely or if the AV should instead perform a safe stop. The currently best modality for in-vehicle warnings can also be optimized based on the situation in the cabin (e.g. not rely on visual HMI when driver is reading or looking at a phone). From the UX perspective, the ride in an AV can be adapted to the state and activity of the driver and passengers. Further, with tracking of face expressions, gestures and body position, the emotional state and response of the driver and/or passengers can be used to evaluate the automated vehicle's actions in traffic. Mapping of all passengers in the AV will enable new methods of understanding how social interaction between passengers, but also between passengers and the intelligent car will look like in the future.

We started with the in-dept analysis of in-vehicle scenarios that are most relevant to safety and UX in AVs of SAE level 2-5. The analysis was performed in two literature review rounds: (i) reports on driver behaviour and potential correlation to accident cases[3, 4, 5, 6, 7, 8] and (ii) literature on most common activities in autonomous vehicles. Since AVs today mainly fall into categories SAE2 and SAE3, the latter is performed by study the surveys of what people want to do in autonomous cars[9, 10] instead of statistics reports.

The list of in-vehicle activities, derived from the literature review, has then been reviewed and categorized by the project team consisting of multidisciplinary researchers, with regards also to the availability of related recognition algorithms.

The initial list of activities are categorized as follows:

- Individual activities
  - Characterization: child/adult, height, age
  - Seat occupancy
  - Object classification per position/seat
  - Body/hand/head position
  - Facial expressions
  - Eye gaze: eyes on road, eye-lid opening, pupil size

*{martin.torstensson, thanh.bui, david.lindstrom, cristofer.englund, boris.duran}@ri.se

[1]The collaboration research project between RISE and Smart Eye, funded by Fordonsstrategisk forskning och innovation (FFI)

  - Forgotten child inside the car
  - Attention and focus, cognitive state: drowsiness, day-dreaming, sleeping
  - Car-passenger interaction: center stack, warning-alert-response, sudden change in behavior as a reaction to something unknown to the car
  - Health state: feeling sick, driving under influence of drugs and alcohol
- Group activities
  - Interaction: talking, fighting, playing, kissing, arguing, gestures, eye contact
  - Body language: approaching behavior
  - Inside-outside the car: gesture
- Unwanted behavior in a shared vehicle
  - Destroying the interior,
  - Violent behavior,
  - Assault,
  - Molest

In addition to the behaviours of interest to be recognized, the following preconditions were also considered in the next steps:

- Input data are video sequences from in-vehicle environment.
- Multiple RGB-IR cameras with combined view covering all vehicle seat positions.
- Multiple persons identified by occupied seat positions.
- Using computer vision and deep learning methods.
- Limitation of computing and memory resources (embedded hardware) and capability to trade-off between accuracy and speed.

From the above initial set of scenarios, we conducted a second literature review round on state-of-the-art of algorithms[11, 12, 13, 14, 15, 16], the publicly available training datasets and pre-trained weights.

We decided to use a module-based hierarchical network architecture, where feature extraction and fusion are at different network phases. Selection of this specific architecture was based on the below identified basic mapping requirements:

- Object detection and recognition
- Seat occupancy and driver/passenger body poses
- Emotions
- Face detection and face landmark
- Individual activities
- Interactions: Person with objects and person-to-person

In the following section, we will only describe the major component of the system: In-cabin activity recognition.

2 IN CABIN ACTIVITY RECOGNITION SYSTEM REALIZATION

2.1 System architecture and components

The proposed scheme of the activity-recognition part of DRAMA system is illustrated in Fig. 1. The figure describes the pipeline of how different mapping classifications are recognized from the input RGB-IR video sequences with the last layer providing activity recognition. The input sequences of images are down-sampled (and converted to gray-scale if captured images are RGB, i.e. during daytime). These prepared images are then fed to the three parallel processing modules: (i) body posture recognition, (ii) object recognition and (iii) optical flows. The body posture recognition provides skeleton model estimations of all persons in the car. The object recognition module provides object classification and related 2D location for objects falling into the classes of interest. Optical flow images capture the local movements of image pixels, image stabilizer and image background subtraction is also considered, to only capture movements highly related to the activities of interest in later recognition steps.

The features extracted from these three modules for each time stamp are then concatenated into longer fixed-length feature vector. This combined feature vector is considered as a snapshot of all important information at the specific time stamp. The feature in turn becomes input to a set of dense neural network layers. The last module in the chain is one or several LSTM [17] layers that will capture the timely orders of instant recognized activities to be able to detect more complex activities.

The RGB or IR images are converted into grayscale for the optical flow feature extractor. During the collection of the images, the different camera feeds were concatenated into one, effectively creating one single video combining

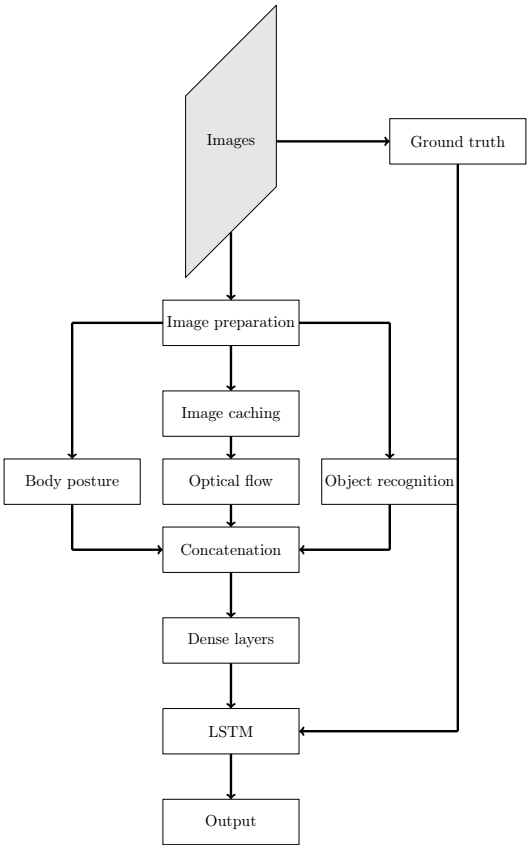

Figure 1: Prototype system architecture

images from all cameras. The intention was to synchronize the cameras to give one large frame for each timestamp. The downside is a risk for duplication and/or dropping of frames, which was regarded less severe in comparison to unsynchronized videos. The network required sequences of images rather than videos. Therefore, the videos were first extracted into frames. Image rotation and alignment techniques were also applied depending on different camera position settings.

The preprocessing is followed by feature extraction which can be divided into three different parts: Body posture recognition, optical flow and object recognition. We selected PoseNet [18] for body posture recognition, which creates multi-person skeletal models based on 2D coordinates of body joints. Another considered alternative is OpenPose [19], which includes hand gesture. The features extracted by optical flow capture the short-term temporal context in the sequence data and is calculated with the Farneback algorithm [20] for all pairs of consecutive images, except the first image. Trade-off of time and accuracy can also be done by selected image pairs of different time-gap scales. The Farneback algorithm is used to create dense optical flow field, which gives two values for each pixel: one for the vertical and one for the horizontal movements. A histogram of the angles created by these vectors is then created in order to increase the robustness. Two different models were selected as candidates for object recognition module: a regions of interest (RoI) based implementation of mobilenetv2 [21] and YOLOv3 [22]. The RoI model is applied to smaller cut out parts of the original image to recognize objects of interest in that given part of the image. Recognized object classes and position will provide valuable input for action recognition that involve objects. This will also provide safety-related information e.g. objects laying around that can cause damage or injury in some driving situations.

The outputs from these feature extractors are concatenated and used to train a combination of dense neural networks and LSTMs. Additionally layers including Batchnorm [23], leaky ReLU [24] and dropout are also used. Concatenation is used to combine features extracted from different modules. By doing this, we have created a modular system where the feature extractors can easily be added, removed or replaced when needed. We will then execute the next series of experiments to validate the performance of different components in the network in the overall architecture by performing step by step adding/removing component feature into the concatenated one. The LSTM complements optical flow feature with time-related futures of higher abstraction level (e.g. a passenger handing over a phone to driver)

2.2 Data capture

There is unfortunately no publicly available dataset for training of driver and passenger in-vehicle activities. In this project, we performed several rounds of data capture and annotation for the designed system. The early rounds of data capture were performed in a simulation environment with limited set of objects and activities

that were enough to preliminary validate the system performance while still maintaining the generality and avoid the potential correlations to any specific types of objects or persons. The later rounds of data capture will be performed in a controlled car environment with data capture facilities from Smart Eye.

References

[1] Jack Stewart. Self-Driving Cars Won't Just Watch the Road. They'll Watch You, Too. *Wired*, February 2017.

[2] Tarek El Dokor. Autonomous Vehicles Need In-Cabin Cameras to Monitor Drivers, October 2016.

[3] Trent W. Victor, Marco Dozza, Jonas Bärgman, Christian-Nils \AAkerberg Boda, Johan Engström, and Gustav Markkula. Analysis of Naturalistic Driving Study Data: Safer Glances, Driver Inattention, and Crash Risk. 2014.

[4] Jeffrey S. Hickman, Richard J. Hanowski, and Joseph Bocanegra. *Distraction in Commercial Trucks and Buses: Assessing Prevalence and Risk in Conjunction with Crashes and Near-Crashes*. 2010.

[5] Thomas A. Dingus, Feng Guo, Suzie Lee, Jonathan F. Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *PNAS*, 113(10):2636–2641, March 2016.

[6] Transportation Research Board, Engineering National Academies of Sciences, and Medicine. *Design of the In-Vehicle Driving Behavior and Crash Risk Study*. The National Academies Press, Washington, DC, 2011.

[7] United States Department of Transportation. Distracted driving and driver, roadway and environmental factors. pages 185–214, 2011.

[8] National Center for Statistics and Analysis. Distracted driving 2016. Research Note DOT HS 812 517, Washington, DC: National Highway Traffic Safety Administration, 2018.

[9] Sofia Jorlöv, Katarina Bohman, and Annika Larsson. Seating Positions and Activities in Highly Automated Cars – A Qualitative Study of Future Automated Driving Scenarios. In *IRCOBI Conference Proceedings*, 2017.

[10] Saptarshi Das, Ashok Sekar, Roger Chen, Hyung Chul Kim, Timothy Wallington, and Eric Williams. Impacts of Autonomous Vehicles on Consumers Time-Use Patterns. *Challenges*, 8:32, 2017.

[11] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. Hierarchical Deep Temporal Models for Group Activity Recognition. *arXiv:1607.02643 [cs]*, July 2016. arXiv: 1607.02643.

[12] A. Iosifidis, A. Tefas, and I. Pitas. Multi-view Human Action Recognition: A Survey. In *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 522–525, October 2013.

[13] Tong Hao, Dan Wu, Qian Wang, and Jin-Sheng Sun. Multi-view representation learning for multi-view action recognition. *Journal of Visual Communication and Image Representation*, 48:453–460, October 2017.

[14] Limin Wang, Yu Qiao, and Xiaoou Tang. Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4305–4314, June 2015. arXiv: 1505.04868.

[15] S. Biswas and J. Gall. Structural Recurrent Neural Network (SRNN) for Group Activity Analysis. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1625–1632, March 2018.

[16] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going Deeper into Action Recognition: A Survey. *arXiv:1605.04988 [cs]*, May 2016. arXiv: 1605.04988.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[18] Ross Wightman. A Python port of Google TensorFlow.js PoseNet (Real-time Human Pose Estimation): rwightman/posenet-python, May 2019. original-date: 2019-01-04T00:34:43Z.

[19] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2d Pose Estimation using Part Affinity Fields. *arXiv:1812.08008 [cs]*, December 2018. arXiv: 1812.08008.

[20] Gunnar Farnebäck. Two-Frame Motion Estimation Based on Polynomial Expansion. In Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Josef Bigun, and Tomas Gustavsson, editors, *Image Analysis*, volume 2749, pages 363–370. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[21] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv:1801.04381 [cs]*, January 2018. arXiv: 1801.04381.

[22] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, April 2018. arXiv: 1804.02767.

[23] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015. event-place: Lille, France.

[24] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.