Intention Recognition for Real-time Automotive 3D situation awareness (IRRA)

Public report

Project within FFI EMK

Author Koen Vellenga
H. Joe Steinhauer
Leon Sütfeld
Svitlana Finér

Date 2024-04-30

FFI Fordonsstrategisk Forskning och Innovation
VINNOVA Energimyndigheten TRAFIKVERKET FKG VOLVO SCANIA VOLVO

Jan 2024

Content

FFI in short

FFI is a partnership between the Swedish government and automotive industry for joint funding of research, innovation and development concentrating on Climate & Environment and Safety. FFI has R&D activities worth approx. €100 million per year, of which about €40 is governmental funding.

For more information: www.vinnova.se/ffi

1. Summary

Intention recognition is the task of inferring an agent's intention based on its previous actions. It is crucial for human social intelligence which in turn enables understanding of, and the ability to predict, other humans' behaviours, such as for example other drivers' intent to overtake, stop, turn, or switch lanes. For making situation-based decisions, both autonomous and human drivers need to take the intentions of surrounding vehicles into account. This is especially true in a mix of autonomous and human drivers.

Existing algorithms and models for intention recognition need to be improved with respect to accuracy, robustness, transparency, and scalability, in order to meet the requirements of the Swedish automotive industry and Trafikverket. It is an open research question how to meet these requirements. This lack of knowledge is a bottleneck for the automotive industry prohibiting the creation of novel advanced and intelligent automotive services and products based on social intelligence and intention recognition.

The overarching goal of the project is knowledge transfer to industry of novel algorithms and models for intention recognition and about their interdependencies with available industrial data and needs, on a level sufficient for enabling ensuing commercial exploitation of the project results. To achieve this, the project will develop new algorithms for intention recognition specifically aimed for the Swedish automotive industry, based on current state-of-the-art in intention recognition and AI, as well as in statistical learning and sensor technology. The main coordinator is Volvo Cars, the industrial partner is SmartEye, and the academic partners are RISE AB and Högskolan i Skövde. Last, Trafikverket is the Swedish government partner.

2. Sammanfattning på svenska

Intention recognition är uppgiften att dra slutsatser om en agents avsikter baserat på dess tidigare handlingar. Det är avgörande för mänsklig social intelligens, vilket i sin tur möjliggör förståelse av och förmågan att förutsäga andra människors beteenden, såsom till exempel andra förarens avsikt att köra om, stanna, svänga eller byta körfält. För att fatta beslut baserade på situationen behöver både autonoma och mänskliga förare ta hänsyn till omgivande fordonens avsikter. Detta gäller särskilt i en blandning av autonoma och mänskliga förare.

Befintliga algoritmer och modeller för avsiktsigenkänning måste förbättras när det gäller noggrannhet, robusthet, öppenhet och skalbarhet för att uppfylla kraven från den svenska fordonsindustrin och Trafikverket. Det är en öppen forskningsfråga hur dessa krav kan uppfyllas. Denna brist på kunskap utgör en flaskhals för fordonsindustrin och hindrar skapandet av nya avancerade och intelligenta fordons- och produkttjänster baserade på social intelligens och avsiktsigenkänning.

Det övergripande målet med projektet är kunskapsöverföring till industrin om nya algoritmer och modeller för avsiktsigenkänning och deras beroenden av tillgängliga industriella data och behov, på en nivå som är tillräcklig för att möjliggöra efterföljande kommersiell exploatering av projektresultaten. För att uppnå detta kommer projektet att utveckla nya algoritmer för avsiktsigenkänning som är speciellt inriktade på den svenska fordonsindustrin, baserat på den aktuella state-of-the-art inom avsiktsigenkänning och artificiell intelligens, samt inom statistisk inlärning och sensor-teknologi.

Huvudansvarig och koordinator är Volvo Cars. Industriell part är SmartEye. Akademiska partners är RISE AB och Högskolan i Skövde. Svensk myndighetspartner är Trafikverket.

3. Background

Problem description

Intention recognition is the task of inferring an agent's intention based on its previous actions. It is fundamental for social intelligence, and it is used for predicting and understanding other humans' behaviour, allowing us to act accordingly. In the automotive setting, such behaviour could be fellow road users' intentions and actions, or it could be the intentions of a group of road users in traffic systems such as a road crossing. It can also be about the vehicle automatically detecting the intentions of the driver or the passengers for automatic activation or configuration of, for example, security systems in the vehicle. In all these examples, for making intelligent situation-based decisions, autonomous and human drivers as well as driver assisting systems need to take the intentions of surrounding vehicles, or drivers, into account.

Intention recognition can be conceived in two very different ways. In the first sense, which has received much attention in the research community, intentions are identified with observable actions. This approach has the advantage that the recognition task can be approached using traditional machine learning inference techniques to predict intentions based on observed sequences of actions. However, this approach overlooks both that intentions can be used for distinguishing between two identical action sequences (Demiris, 2007), and that intentions often are independent of observations of some particular actions (Meltzoff, 2007).

In the second sense, by contrast, intentions are conceived as distinct entities that, instead of being directly observable, have to be inferred from observed actions (and possibly a background theory), using for example temporal, causal, or spatial relations between observed actions. While rendering the recognition task exponentially more complex from a computational point of view, this sense of intentions addresses the two objections to the first sense above and is more in line with how intentions are conceived.

The problem we address is that algorithms for intention recognition in the first sense enable creation of services that merely detect or predict actions, while not making it possible to reveal the intention, in the second sense, behind observed actions. For example, algorithms for intentions in the first sense can tell that a car has stopped or will stop, but they cannot tell why it stopped or will stop.

In order to build capability for creating automotive services based on intention recognition in the second sense, and because of the higher complexity, existing algorithms need to be improved with respect to speed, precision, transparency, robustness, and scalability. For example, four issues that need improvement are:

- Machine learning algorithms developed for the first sense of intention recognition lack precision and coverage when applied for intention recognition of the second sense.
- Algorithms for the second sense of intention recognition rely on discernibility between observed sequences of actions for inferring intentions, which contributes to the complexity and slows down inference.
- Logic-based approaches are not robust to uncertainty and do not scale well, but they are transparent and can have speedy inference in special cases.
- Statistical learning approaches are not transparent and require extensive learning for achieving precision, but they are robust to uncertainty and very fast in inference.

To address the problem, IRRA set out to use state-of-the-art knowledge and technology to improve existing algorithms for intention recognition and subsequently verify the suitability of the developed solutions in a number of proof-of-concept implementations based on industrial specifications. We focus on two areas for improvements: i) the area of *action recognition* where

the observable actions are extracted from sensor data; and ii) the area of *inference* of intentions from recognized actions. This leads to the following two research objectives:

1. Improve existing algorithms for action recognition to increase discernibility of observed sequences of action through computer vision and novel sensor technology, and through utilizing novel technology for localization of vehicles.
2. Improve existing algorithms for inference of intentions based on observed sequences of actions. For example: using (i) self-supervised reinforcement learning for improving machine learning based approaches; (ii) data-driven reasoning (Prade, 2016) for improving logic-based approaches; and (iii) novel result in statistics concerning causality (Pearl, 2009) to improve approaches based on statistical learning.

The specific requirements for the new algorithms are identified through concrete use cases based on three scenarios, described in Section *Project Contents* below. The scenarios reflect three industrially relevant perspectives of intention recognition: the *driver perspective*, the *vehicle perspective*, and the *systems perspective*. The use cases will be developed in the beginning of the project.

State-of-the-Art ()

The task of intention recognition is performed in two steps: *action recognition* and *inference of intentions.* Action recognition is the basic task where the agent's actions are recognized. This can be made through targeted processing of video data (or generally: sensor data) for enabling the detection of a given set of relevant actions. IRRA set out to improve the action recognition step by fusing a number of sensor sources. In particular, wemainly use video to recognize actions and the recognized actions will serve as input to the inference step, together with localization, and driver state. The inference step consists of matching the discovered action sequences with a model that associates action sequences with intentions. The inference step involves also the selection of one intention among a potentially large number of candidates. Next, we give a short state of the art for the main research topics of the proposed project: video-analysis, maps for localization, and intention recognition.

Video analysis for action recognition

A well-studied field in computer vision is *scene understanding* with methods suitable for action recognition. The procedure of scene understanding can be broadly divided to semantic segmentation and scene classification. Semantic segmentation is the first step towards scene understanding. It is based on low-level features, for instance colour, edges, and illumination. The approaches for feature selection have been conveyed also in the subtasks of lane and road detection, traffic sign recognition, and vehicle detection. Furthermore, contextual information is important for semantic segmentation. Using contextual information, the generally applied models of scene understanding can be subdivided as graphical models, convolutional networks, cascaded classifiers, and edge detection-based approaches (Seyedhosseini et al., 2016).

Markov Random Fields (MRF) and Conditional Random Fields (CRF) are the most popular Graphical Models approaches. In Pele et al. (2009), an inference technique was proposed for MRF to minimize a unified energy function. In Xuming (2004), a CRF method was presented for labelling images. In (Sudderth et al., 2008), a hierarchical Dirichlet process was developed to model visual scenes. Convolutional networks-based approaches have emerged recently to dominate most of the state-of-the-art. In (Grangier et al., 2009), a convolutional network was trained for scene parsing. In (Chen et al., 2014) a deep convolutional network in combination with CRF was shown to improve the semantic segmentation performance.

In Heitz et al. (2009) a different architecture for making multiple classifiers fusion into a cascaded classifier model was proposed for scene understanding. In Li et al. (2012) a feedback enabled cascaded classification model was presented to mutually optimize several subtasks in scene understanding where each classifier was considered in series. Therefore, the training process of a cascaded classifier model was considerably simpler than convolutional networks (Seyedhosseini et al., 2016). Edge Detection based approaches are also popular. Various unsupervised methods have been applied for edge detection. In Seyedhosseini et al. (2016) a contextual hierarchical model was used to distinguish between "patches centered on an edge pixel" and "patches centered on a non-edge pixel".

As most scenes are composed of objects in a highly variable layout, scene classification is an important problem for surroundings perception. In the literature, scene classification has been focused on binary problems, such as distinguishing indoor from outdoor scenes. Inspired by the way how the human perception system works, numerous efforts have been dedicated to classifying a large number of scene categories. The most popular method is the "bag-of-features". It represents images as order-less collections. The features can be extracted by the Scale-invariant feature transform (SIFT) (Lowe, 2004) or the HOG method (Dalal et al., 2004). In Sánchez et al. (2013), the SIFT was presented to abstract visual features and these features were encoded to a Fisher kernel framework for scene classification. In Zhou et al. (2014), a convolutional neural network (CNN) was proposed to perform scene classification.

Recently, the model of "bag-of-semantics" was presented. In this model, an image is extracted as a semantic feature space. It has the capability to perform a spatially localized semantic mapping. In Su et al. (2012) a number of classifiers for each individual semantic class was trained for object classification. In Li et al. (2014), a high-level image representation encoding object appearance and spatial location information was presented. In Dixit et al. (2015) a semantic Fisher vector, which is an extension of the Fisher vector to bag-of-semantics, was applied to classify image patches.

Intention recognition

A model for intention recognition typically uses so-called plan-libraries, relating sequences of actions with intentions. Plan-libraries can be man-made or induced through machine learning. Recognition of intentions is made by inference, based on the plan-library and an observed sequence of actions. Observations of sequences of actions are made via action recognition techniques, as described above. For inference of intentions, one needs first to match the observed actions with the action sequences modelled in the plan-library, and then to select an intention based on the matched sequence. It is the two steps in the inference phase that give rise to the computational complexity of, and challenges in modelling, intention recognition. Among the challenging factors are domain-specific characteristics such as adversary agents disguising their intention (Mao and Gratch, 2004); concurrent intentions (Hu and Yung, 2008); multiple agents (Sadelik and Kautz, 2010); uncertainty about intentions (Charniak and Goldman, 1993); partial observability of actions; and noise in agents' plan execution (e.g., Roy et al., 2007), in monitoring patients with Alzheimer's disease). Consequently, one of the main foci in intention recognition has been on finding models bridging this complexity, for example, by exploiting hierarchical structures in plans (Kautz and Allen, 1986).

The formalization of intention recognition was initially based on methods from logic (Schmidt 1978; Kautz and Allen 1986) and on AI-approaches to planning (Kautz, 1987). Hobbs et al. (1993) were among the first to use abduction to reason about intent, which has become the predominant approach in the logic-based community. Charniak and Goldman (1993) introduced a

probabilistic approach to the problem, based on Bayesian Belief Networks, which is also the approach that has received the most attention. Both the logical and Bayesian approaches are still actively researched (Sadri, 2010).

There is a long tradition of using statistical learning for intention recognition; however, these approaches often identify intent with directly observable actions and cannot feasibly and efficiently be used for recognizing intention based on relations between observed actions (as we aim for in the proposed project). For example, Bui et al. (2002) used Hidden Markov Models to infer intentions as hidden states, and Schultz et al. (2015) use a variant of Conditional Random Fields. Many of the more successful approaches have been hybrids of logical- and probability-based approaches. Goldman et al. (1999) use probabilistic Horn abduction, and Huber et al. (1994) use Bayesian Logic Programs.

Intention recognition has been applied in several diverse domains, just to mention a few: automotive (Schultz et al., 2015) for recognizing pedestrian intent, and (Berndt et al., 2008) for early detection of driver intention, and (Da Lio et al., 2015) for automated co-drivers; elderly care (Roy 2007) for monitoring Alzheimer patients; military (Mao and Gratch, 2004) for understanding adversary actions; and team sports (Sadelik and Kautz 2010; Intille 1999; Gudmunsson and Horton 2017).

4. Purpose, research questions and method

Project contents

The proposed project will be based on work with a number use cases from three scenarios. The project will initially define these use cases based on available data and concrete exploitation possibilities. The scenarios cover three aspects of intention recognition: in-cabin driver intentions; the intentions of drivers in vehicles close to host vehicle; and intentions of the participants in a traffic system.

- **Scenario 1 (In-vehicle and driver environment perspective: driver intention and state modelling)**. Driver intention can have decisive impact on the utility of the activation of various vehicle systems (e.g., safety or automation). The project will initially develop use cases that reflect concrete needs of driver intention recognition from the automotive industry and/or the society.

- **Scenario 2 (Vehicle perspective: Intentions of drivers of interacting vehicles)** An important area of research is that of vehicle intention recognition, where the host vehicle can benefit from understanding and predicting the intention of vehicles that affect the host vehicle's range of action alternatives. The project will in the initial phases develop concrete use cases for vehicle intention recognition reflecting needs from the automotive industry and/or the society.

- **Scenario 3 (System-based intention recognition).** System-level intention recognition is characterized by understanding an individual person's or vehicle's behaviour in the context of the traffic state around them, as opposed to focusing merely on its own past behaviour. The system-level approach is characterized by (1) the use of V2X technology to gain a more complete picture of the traffic state than is possible from a single point of view, (2) the notion that the system state determines behavioural priors and that taking these priors into account can improve the accuracy of predicted actions, and (3) the objective of improving the safety and efficiency for all affected parties through understanding and managing collective behaviour and dynamics of traffic.

Methodology

This section aims to provide a high-level outline of the methodological underpinnings of the work carried out in the project. To enable the data collection, a collaboration between Volvo Cars and Smart Eye AB was required to integrate the systems into a vehicle. The driver and environment monitoring work package (WP-2) relied on open-source datasets, previously collected datasets, as well as the data collected from our own data collection endeavours. The exploratory analyses of how and what data to analyse, served as a foundation for developing the intention recognition algorithms in the other work packages. The "Vehicle perspective: Intentions of drivers of interacting vehicles" (WP-3) research work was based on the expertise from senior researchers at RISE. Work package 4 (System perspective of intentions) outlines future work based on the encountered challenges by reviewing existing methodologies and datasets. Work package 5, Intention Recognition for multi-agent scenarios, is jointly carried out by Högskolan i Skövde and Volvo Cars. For the industrial PhD, weekly meetings are held to supervise and align the research. The dissemination and demonstration of the work is conducted through scientific publications, and workshops with the project partners.

5. Objective

The project aims at developing technology that enables introduction of new services and products that build on intention recognition. Some of the areas that may benefit from this are passive/active safety, autonomous drive, and energy usage prediction. In particular, the present proposal fit the objectives of the subarea *Green, Safe, Autonomous & Connected functionalities* of the programme.

This project aimed to contribute to the following overarching FFI targets:

- Increased capacity for research and innovation ensuring the competitiveness of the Swedish automotive industry and Trafikverket.
- Encouraged collaboration between the industrial, academic, and societal sectors.

The proposed project will contribute to the following targets of the program for Elektronik, Mjukvara, och Kommunikation:

1. The project will *contribute to the competitiveness of Swedish automotive industry* through exploration and development of methods for industrial exploitation of artificial intelligence.
2. The project will *reinforce collaboration between industry, research institut*es, *academia*, and *government agencies*. All four sectors are represented by the project partners. The sectors represent different concept of values and modes of operation, and the proposed project will lead to improved practices for similar future collaborations.
3. Among the applications that can benefit greatly from the results of IRRA, there are many that also fit into the area of Green, Safe, Autonomous & Connected functionalities. In particular, the project will initially develop concrete use cases guiding the research, and three potential targets for this development are autonomous drive, safety and user prediction for function- and service automation.

While the aim of the proposed project is an enabling technology, we would like to point out that the proposed project is also relevant for the programs Trafiksäkerhet och automatiserade fordon and Effektiva och uppkopplade transportsystem.

6. Results and deliverables

6.1 Data collection and procurement (WP 1)

Work Package 1	Data collection and procurement
Responsible	Volvo Cars
Other participants	TrV, SEYE

1. Data collection

For this project data was collected in a sensor equipped vehicle (see Figure 1. Data collection vehicle with decal informing road users of the collection process.) on real roads (see general region in Figure 2).

Figure 1. Data collection vehicle with decal informing road users of the collection process.

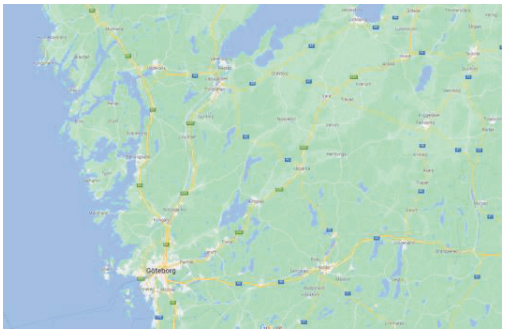

Figure 2. Overview of the data collection region

To capture the exterior driving scene, two full HD USB cameras were used, mounted inside of the car, one facing forward and one facing backwards. The decal on the side of the car informed road users that the data collection process was ongoing and if anyone did not wish to be part of the research data, they could communicate it via a specially created webpage.

To capture the interior of the vehicle including the driver, two Smart Eye camera systems were used. One-camera driver monitoring system (DMS, see Figure 3a) was placed behind the steering wheel capturing head, face and eyes of the driver, while one-camera cabin monitoring system (CMS see Figure 3b) was placed above the rear-view mirror capturing the entire cabin including full upper body of the driver (see Figure 4). Both camera systems operate using near IR light that makes them independent from the ambient light, which allows collecting data both when it is light and dark outside. Recorded signals include, but are not limited to, head position in 3D, head orientation, face tracking, gaze direction, glance area estimation, eye tracking, eye lid opening, pupil tracking, mouth opening, body pose, objects in the car, objects in hands, and others. Most of the signals contain information about the signal quality, which can be further used to calculate uncertainty of the intention recognition output (see Figure 5 and Figure 6).

Figure 3. (a) Smart Eye DMS system to the left and (b) prototype CMS system to the right.

Figure 4. Smart Eye DMS and CMS installation in the data collection vehicle.

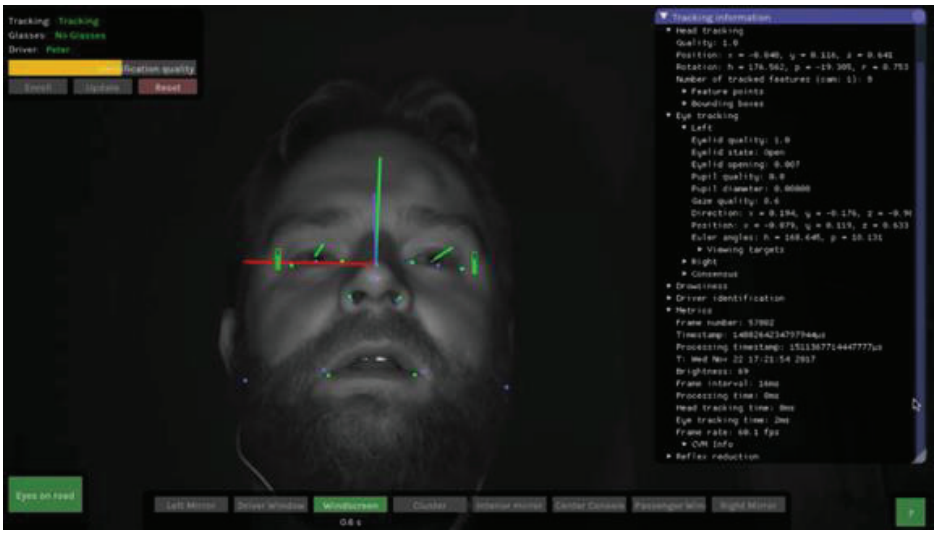

Figure 5. Example of Smart Eye DMS image capture and signals.

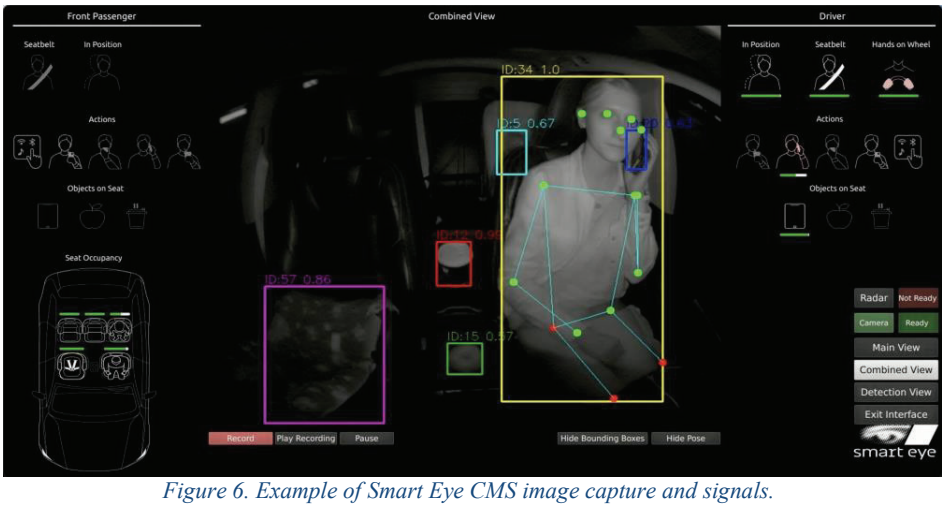

Figure 6. Example of Smart Eye CMS image capture and signals.

To capture vehicle dynamics an internal data collection unit was used. It captured signals, such as yaw-rate, steering wheel angle, velocity. Additionally, it synchronized the data from the various streams, to ensure that all observations had a timestamp, and sent the data via a safe connection back to the secured portals.

Data collection process on the high level is presented in Figure 7. After the data had been collected, the raw video footage was only retained for four weeks, due to privacy regulations, and could only be accessed by verified project members in a contained room. There the data was processed and harmonized. After that the processed data excluding any image and video footage was securely stored with limited access for verified project members.

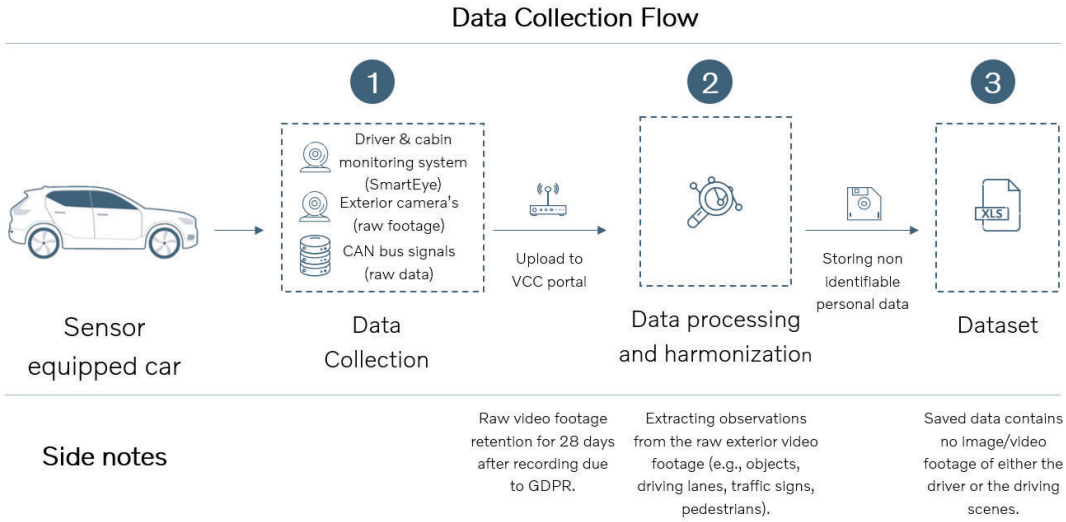

Figure 7. Overview of the data collection process for this research project.

Reflections and challenges. Transferring gigabytes of video data over a secured wireless mobile network connection has proven to be difficult. As a result, we lost hours of driving data

due to connection and transfer errors. Moreover, handling, synchronizing, and labeling of the data requires a lot of time and resources.

2. Data procurement

Any algorithm development and testing require data. And neural network algorithms require lots of data. Additionally, in most cases the data has to be labeled and contain specific use cases. In this project data was collected in the real car on the real roads, but due to limitations in time and resources that data could not cover all the algorithm development that was planned and conducted in this project. Therefore, in addition to that data partners looked for other data sources. Some partners found suitable data internally, others turned to open-source datasets. In some cases, additional annotations and labels were needed and in others they all were already there. The great efforts of the team resulted in suitable datasets covering all project needs.

6.2 Driver and environment monitoring (WP 2)

Work Package 2	Driver and environment monitoring
Responsible	RISE
Other participants	Volvo Cars, SEYE

6.2.1 Object detection, depth and pose estimation exploration.

In order to detect the objects and traffic entities in front and around the car, and estimate their spatial location, we implemented a system that operates on dashcam video streams. To achieve a high framerate, we used the at-the-time state-of-the-art YOLOv5 to detect entities. YOLO provides bounding boxes and labels for the detected classes of objects. In order to find the distance of the objects, we used another, pretrained neural network model that computes depth from video. It was trained on data from the KITTI dataset which consists of video, laser scanner, and human tagged entities like cars, cyclists, vegetation, traffic signs/poles, pedestrians, sidewalks, etc. Figure 8 shows an interface that visualizes the application of the two neural network models on a video stream.

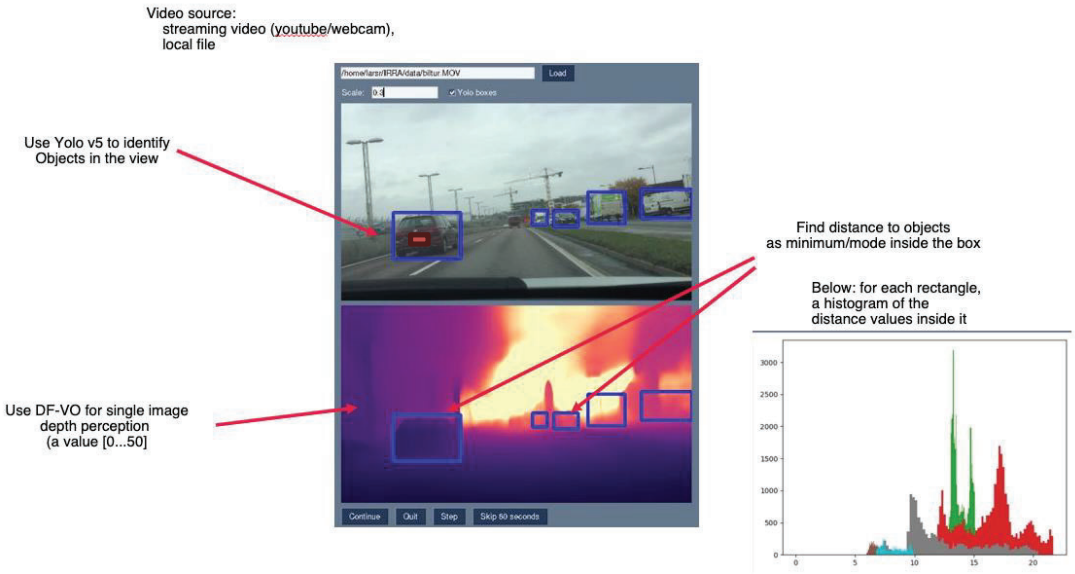

Figure 8. A Vision-to-Distance pre-trained neural network for determining the distances to objects. The histogram to the right shows the distributions of per-pixel distances in the image's bounding boxes as detected by the model (colour coded for the five bounding boxes). The histogram lets us determine the distance of the nearest object to be around 6-7 meters.

	cls	frame	source	x0	x1	y0	y1
0	car	0	JAAD_clips/video_0333.mp4	530	560	650	679
1	car	0	JAAD_clips/video_0333.mp4	1010	1060	685	712
2	traffic light	0	JAAD_clips/video_0333.mp4	968	990	591	622
3	car	0	JAAD_clips/video_0333.mp4	959	1010	679	705
4	car	0	JAAD_clips/video_0333.mp4	928	997	680	706
...	...	...	...	...	...	...	...
1027358	person	239	JAAD_clips/video_0268.mp4	1535	1563	672	751
1027359	truck	239	JAAD_clips/video_0268.mp4	674	780	661	770
1027360	bus	239	JAAD_clips/video_0268.mp4	1017	1240	602	824
1027361	car	239	JAAD_clips/video_0268.mp4	838	927	708	776
1027362	person	239	JAAD_clips/video_0268.mp4	253	483	649	1039

[1027363 rows x 7 columns]

Figure 9. Labels and locations of objects found in a sequence of video stream frames.

When the software operates on a continuous video stream, it generates a table of rows records with object labels and positions, see figure 9. The figure shows the software applied to the videos from the JAAD dataset.

We have also investigated the detection of cars and pedestrians using YOLO4 and various lightweight versions of YOLO. The primary goal was to measure how often the algorithm missed detecting a vehicle or a pedestrian. In addition, our exploration extended to lane and blinker detection. Furthermore, we extracted locations of joints (such as arms, legs, and head direction) of detected individuals using Alpha Pose, a multi-person pose estimation system, along with Pose Flow as a tracking algorithm. Figure 10 illustrates the accurate estimation of multi-person poses. However, occasional misjudgements of human-like objects and the neglect of small human objects are among the weaknesses of this algorithm.

Figure 10. Multi pose estimation using Alpha Pose.

The output of this stage is intended to be used as input to the more high-level, and semantically interpreted description of the scene.

While all of the subcomponents of this design have been independently improved in the opensource community in the last few years, the overall design of the component-based vision system is still valid, and can easily and seamlessly benefit from the speed improvements by just upgrading.

6.2.2 Data processing and automatic environment representations

In addition to exploration of object detection, and proximity estimation of road users, we rely on an automated pipeline to process the videos collected with the data collection vehicle. Similar to the exploration above, we implemented a pipeline with YOLOv8 and ByteTrack (Zhang et al., 2022) to identify road users across different frames. To segment the drivable area in the frames, we rely on Grounded-Dino (Liu et al., 2023) and Grounded-SAM, which enables segmenting any arbitrary object based on a text prompt (see Figure 11). Afterwards, the tracking results and bounding boxes are used to construct dynamic discrete evolving graphs for each video frame (Figure 12).

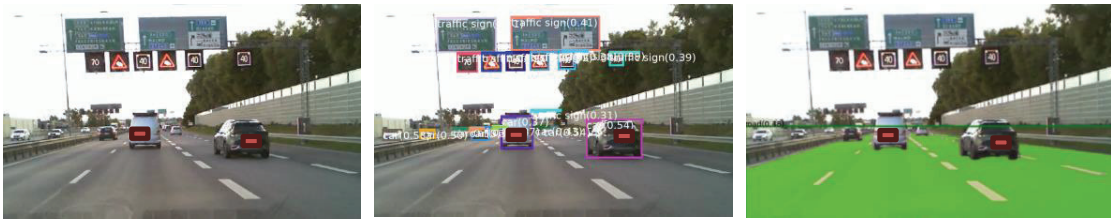

Figure 11. From left to right: Original observation, object detection, drivable road segmentation.

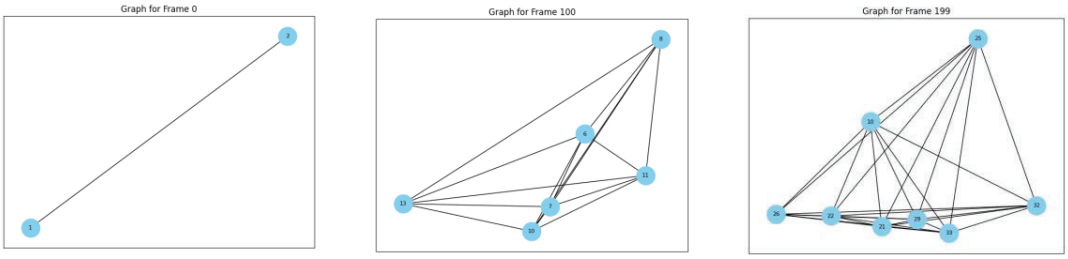

Figure 12. Representation of the driving scene as a connected graph for various frames. Each node represents a unique tracking ID of an observed object.

6.2.3 – Driver state recognition

Driver state recognition using cameras supports the decision process in the vehicle safety system. Driver states play a crucial role in traffic safety, as they directly influence a driver's ability to respond to road conditions, make decisions, and operate a vehicle safely. Understanding and monitoring these states can significantly reduce the risk of accidents. Here are some key driver states of interest for traffic safety.

Alertness and Attention:
Fully Alert: The ideal state for driving, where the driver is fully engaged, aware of their surroundings, and capable of responding quickly to changes.
Inattentive: Occurs when a driver's attention is diverted from the driving task, whether due to external distractions (e.g., mobile phones, infotainment systems) or internal distractions (e.g., daydreaming).

Drowsiness:
Characterized by reduced alertness and slowed reaction times, often resulting from insufficient sleep, long periods of driving, or monotony. It significantly increases the risk of accidents due to the potential for microsleeps or decreased vigilance.
Fatigued: Beyond just sleepiness, fatigue encompasses overall physical and/or mental exhaustion, impairing the driver's ability to concentrate and make decisions.

Emotional States:
Stressed or Anxious: High levels of stress or anxiety can impair decision-making abilities, increase aggressiveness, and lead to risky driving behaviours.
Angry or Aggressive: Also known as "road rage," this state can result in aggressive driving behaviours, such as speeding, tailgating, and unsafe lane changes.

Impairment:
Alcohol or Drug Impairment: Consumption of alcohol or other drugs significantly impairs judgment, coordination, and reaction times, making driving highly dangerous.
Medication Effects: Certain medications can cause drowsiness, dizziness, or other side effects that impair driving ability.

Cognitive Overload:
Overwhelmed: Occurs when a driver is presented with too much information or too many tasks at once, leading to a decrease in the ability to effectively process information and make safe driving decisions.

Health-Related Impairments:
Vision Impairments: Any condition that affects visual acuity, peripheral vision, depth perception, or colour recognition can compromise driving safety.
Physical Disabilities: Physical limitations can affect a driver's ability to control the vehicle, necessitating specialized adaptations for safe driving.
Medical Conditions: Conditions such as epilepsy, diabetes, or heart disease can pose.

Detection of driver state
Advancements in vehicle technology and driver monitoring systems are increasingly capable of detecting these states in real-time, offering added information in the task of predicting a potential accident.

In the IRRA project Smart Eye focused on providing Alertness and attention on road and handling food together with development of a new variant of drowsiness estimation based on neural networks as described below.

Alertness to road is an important factor as non-alertness leads to slow or no reaction on a forward threat. Eating or drinking degrades the reaction performance as hesitation will be part of the process- where can I put my food? Drowsiness provides the vehicle with added information on reaction time and potential risk for sleep. This part is a very challenging area, and this is where Smart Eye invested most resources.

Driver state drowsiness development is a gradual process that can be monitored through various lower-level signals, primarily focusing on eye movement patterns, pupil dynamics, and other physiological and behavioural cues. Understanding these signals is crucial for developing systems

that can detect and mitigate the risks associated with drowsy driving. Here's a detailed breakdown of how drowsiness can be assessed based on these lower-level signals provided by the driver monitoring system.

Eye opening and closure pattern is a widely used metric for drowsiness detection. It measures the eye opening over time as drowsy drivers tend to have slower blink rates and longer blink durations, leading to increased periods of eye closure. The pupil movements and diameter are another important indicator affected by drowsiness.
Pupil size fluctuations: The diameter of the pupil can vary due to changes in lighting, cognitive load, and drowsiness. Under conditions of constant lighting, variations in pupil size can indicate shifts in alertness, with drowsiness often leading to a reduction in pupil diameter.
Blink frequency: An increase or decrease in blink rate can indicate fatigue. Typically, as a person becomes drowsier, the rate at which they blink may initially increase, followed by a more significant slowdown.
Blink duration: Drowsy individuals often exhibit longer blink durations. Prolonged closures can significantly impair driving performance by increasing reaction times and reducing situational awareness.
Slow Eye Movements (SEMs): As drowsiness sets in, the smooth pursuit movements of the eye can become slower and may include microsleeps, where the eyes may drift slowly off target.
Saccadic Movements: The speed and accuracy of rapid eye movements from one point to another can be affected by drowsiness. A delay in initiating saccades or a decrease in saccadic velocity may indicate fatigue.
Head movement and posture is also an indication where monitoring the position and movement of the head can provide additional cues about a driver's state of alertness. For example, frequent nodding or a drooping head position can indicate drowsiness. Changes in how a driver sits or maintains their posture can also be indicative of fatigue, with slumping or adjustments in seating position being potential signs.
Integration Driver monitoring systems can also potentially analyse facial expressions for signs of fatigue, such as yawning, frequent rubbing of the eyes, or other facial cues indicating tiredness. Typically, an increase in yawning frequency is correlated to drowsiness.

The key to effective drowsiness detection lies in the continuous and real-time analysis of these lower-level signals, allowing for early detection and timely intervention. Such systems are becoming increasingly sophisticated, incorporating not only the direct physiological and behavioural indicators of drowsiness but also contextual factors such as time of day, driving duration, and patterns of vehicle control inputs.
The development of driver drowsiness in the IRRA project is linked to the Karolinska Sleepiness Scale (KSS). Smart Eye integrated above-described signals and trained a machine learning models to assess the driver's state of alertness. By analysing the combination of eye movements, blink patterns, pupil dynamics, and head movements, the system was able to predict an estimated drowsiness level following the KSS scale giving the safety system the ability to take into account the drowsiness level of the driver to provide the right countermeasure like alerting the driver or initiate an earlier countermeasure to prevent an accident.
The development in the project followed standard practices.

Integrating signals to assess drowsiness levels based on the Karolinska Sleepiness Scale (KSS) through a neural network approach involves a sophisticated analysis that combines various physiological and behavioural cues. The KSS is a subjective scale used to measure an individual's level of sleepiness during certain duration (usually 5 min), typically ranging from 1 (extremely

alert) to 9 (very sleepy, great effort to keep awake, fighting sleep). Translating this subjective scale into an objective measure using driver monitoring signals and neural networks involves several steps. The actual data used for the development was pre-collected and pre-annotated.
Neural Network Design: A neural network architecture suitable for time-series analysis and pattern recognition, such as Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs). Long Short-Term Memory (LSTM) networks, a type of RNN, are particularly effective for sequences of data and may be well-suited for detecting patterns in physiological signals over time. After selection of method next is the design of the input layer to accept the extracted features from the preprocessing step. The network needs to process sequences of data to account for the temporal dynamics of drowsiness development. The output layer of the networks is trained to correspond to an estimated KSS level based on the sequence of data that has been provided in the last 5-minute interval. In the end it was formulated as a regression problem (predicting the exact KSS rating) and a classification problem (categorizing the state into bins such as alert, slightly sleepy, very sleepy, etc.). Both methods were tested.

In the validation step cross-validation was introduced to ensure that the model generalizes to unseen data. This involves dividing the collected data into training and validation sets and iteratively training the model while monitoring its performance on the validation set to prevent overfitting.

To conclude, our findings show that by leveraging neural networks to analyse and integrate various physiological and behavioural signals, it's possible to develop a dynamic and accurate system for real-time drowsiness detection based on estimated KSS levels. This approach enables a nuanced understanding of drowsiness development, providing a valuable tool for enhancing driver safety.

6.3 Vehicle perspective: Intentions of drivers of interacting vehicles (WP 3)

Work Package 3	Vehicle perspective: Intentions of drivers of interacting vehicles
Responsible	RISE
Other participants	

In this work package we have explored two different approaches for intention recognition: one logic-based and one statistical model based.

1. Logic-based approach

Traditional logic approaches to intent recognition have been based on automated planning frameworks (Sadri, 2011). In automated planning, one attempts to automatically derive a series of actions that transform an initial state to a given goal state. In the logical approach to intention recognition, the problem is the reverse given a sequence of actions and states of the world, the task is to find out which plan or which goal the actions are aimed at.

We tried a different logical approach to intent recognition by considering intent as a mental state and using an existing logical calculus in a novel way. The existing calculus has been used in legal and normative reasoning and it allows to rule out certain states of affairs (such as whether it is permitted or not to perform an action), based on observations (Sergot, 2001).

We used the calculus and a modal logic, with an operator for intent, to exhaustively generate all possible intentions a driver may have, with respect to a specific driving behaviour, in a traffic situation (such as turning left in a junction.)

The greatest challenge in the task was to transform traffic and vehicle data into valuations of logical atoms with which to, as a conclusion of a logical deduction, logically filter out possible intent of the driver that are inconsistent with observation. One aspect of the challenge was to handle uncertainty and artefacts in the data stream, to which logical models are sensitive. This was accomplished by applying Gaussian smoothing filters to the data, which smoothed out spikes and artefacts. The second aspect of the challenge was the interfacing between data and logical formulas. This was tackled by imposing reasonable (and challengeable) assumptions on driver behaviour.

The approach was implemented in Python and demonstrated to the consortium. The implementation includes the logical calculus, statistical smoothing, and the interface between data and logical evaluations. Based on streaming data of vehicle position and speed the implemented calculus returned a logical encoding of driver intent.

2. Statistical model approach

The approach here is to use a statistical model, Prototype-based classification (Gillblad et al., 2008), which has previously been used for example for troubleshooting technical systems. This is a hybrid between a statistical machine learning model and a knowledge-based representation of the classes. In this case it means that the expected appearance of different potential intentions is described in terms of high-level features, and these human-provided intention descriptions, so called “prototypes”, are then used as training data in a highly regularized statistical machine learning model based on Bayesian statistics. The high-level features are primarily meant to be detected in images around the car, such as surrounding vehicles, pedestrians or other traffic users, traffic lights and signs, car blinkers and break lights, animals, obstacles, etc. Some information may be easier to get from other sources than image analysis. For example, type of road, number of lanes, speed limits, position of traffic lights and pedestrian crossings might be received with higher precision and reliability from a traffic map service than from the image.

The rationale for this approach is (as previously discussed) the high complexity of the intention recognition task together with the very limited amount of real (labelled) data which would be available for each of the interesting intentions we wish to detect. For example, a deep-learning approach going from image to intention in one step, would need a rich variety of cases of each intention to be able to discern what the relevant characteristics of each intention are. Instead, we let the image analysis identify common objects and their properties, and then we use human knowledge to characterize the interesting intentions in terms of these objects and their relations.

Once a system like this will be employed at a larger scale all over the world, sufficient training data will eventually become available. As a steppingstone towards that situation, this knowledge/data hybrid approach is necessary. An advantage with the proposed solution is that it can seamlessly combine human provided prototypes with real data. The prototype is used as an initial prior which is then gradually refined as real data arrives.

The use case we have considered here as illustration of the technique is that there is a car standing still or running slowly in the same lane in front of the own car. To be able to decide whether to overtake it or not, it is important to figure out why it is doing that. E.g., if its intention is to wait for a passenger we might overtake it, but if it is waiting for a red traffic light we should better not.

To this end we have produced a set of prototypes, characterizing different situations and intentions which might explain the behaviour of the car in front of us. Each prototype is associated with a recommendation whether to overtake (if considered safe based on other traffic) or wait behind it (until the situation changes or new evidence is available). That is, in a situation where our car catches up a car ahead in our lane and given a set of features identified in the camera images and from map services, these prototypes can be used to produce a recommendation to overtake or stay behind the other car.

6.4 System perspective of intentions (WP 4)

Work Package 4	System perspective of intentions
Responsible	RISE
Other participants	(Volvo Cars), TrV

Work package 4 was restructured during the runtime of the project. Its new aim was to outline and develop the basis for a successor project to IRRA with a focus on the system perspective of intentions, which otherwise hasn't been actively covered in the project. To this end, a work group around members of RISE and Trafikverket built a consortium and developed the core idea for a project called "Radar-Based Intelligent Traffic State Analysis and Predictions for V2X (RABITS)". The details and state of the RABITS research initiative at the time of writing are presented below after a detailed definition of system-level intention recognition.

We define system-level intention recognition to be characterized by the following 3 aspects:

(1) **Use of V2X information**: Using V2X information to enrich the information used locally on the vehicles to more accurately predict intentions and future behaviour of other road users. By leveraging data from vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications, the system gains a more comprehensive understanding of the traffic environment beyond the immediate sensory range of individual vehicles.

(2) **Behavioural analysis based on system state**: This entails the analysis of behavioural probabilities or priors for behaviour that depend on the state of the traffic system. The configuration (e.g., type, position, speed, signaling, etc.) of vehicles ahead of (or around) a target vehicle can inform behavioural priors for this driver/vehicle. The approach of system-level IR takes such factors into account to improve the overall accuracy of inferred intentions and behavioural predictions for individual vehicles.

(3) **The objective of enhancing safety and efficiency**: System-level intention recognition aims to optimize the flow and efficiency of traffic at a system-wide level, anticipate and mitigate traffic congestions, and improve overall safety. It isn't merely focused on the intentions of individual road users, but about understanding and managing the collective behaviour and dynamics of traffic. It may therefore include recommendations for the behaviour of vehicles that optimize not just an individual's safety and efficiency but the entire traffic system's safety and efficiency.

We illustrate this concept of system-level intention recognition with an example: Sensors installed as part of the infrastructure (e.g., radar sensors on a gantry above the highway) inform a locally run traffic model and AI-based recommendation system which informs approaching vehicles about the traffic state (flow and speed, vehicle constellations, risk-levels, etc.) in the upcoming stretch of road and prompts automated vehicles to, e.g., reduce their speed, increase the distance to the vehicles ahead, temporarily avoid lane changes, or directly assigns lanes to individual vehicles. This could mitigate the development of traffic jams, reduce risk of accidents around congestion and traffic jams, and lead to optimizations in energy use of vehicles in the given area.

Consortium Members
Leon Sütfeld (RISE), Joakim Rosell (RISE), Sepideh Pashami (RISE), Xiaoliang Ma (KTH), Tomas Julner (Trafikverket), Ellen Grummert (VTI)

Project Outline
The RABITS project is planned as a successor to the IRRA project and aims to exploit available radar data in the context of V2X features for increased safety and efficiency on large roads. Potentially extracted features encompass traffic state prediction, intention recognition, anomaly detection, and more. The developed technology is intended (1) for use in advanced ADAS and AD functions in vehicles, either as unidirectional information from the infrastructure or in a collective perception framework, (2) as marketable service features for roadside radar system operators, and (3) to advance the SOTA in traffic analysis and driving behaviour research.

The project has access to a large amount of data from gantry-mounted radar stations across Sweden. The in- tended data collection will happen from ~5-10 stations depending on the use cases (to be determined), likely including the E4 South near Stockholm. All data will be collected, stored and provided by Trafikverket from NavTech Radar sensors. The data features trajectories and speeds of vehicles in a range of approximately 500m per station, is clocked at 4Hz, and includes lane classification.

Use Cases
The focus use cases for this research project will be determined by and in discussion with the use case providers, e.g., Volvo Cars (sought) and NavTech Radar. Potential V2X use cases are ample and encompass safety and efficiency, as well as traffic patterns and driving behaviour:

Safety
◦ Road condition analysis (behavioural imprints of rainy/icy conditions)
◦ Anomalous behaviour detection (e.g., intoxicated drivers, road rage, tailgating; accident risk)
◦ Anomalous traffic flow conditions (e.g., accident detection)
◦ Gap creation (merging at on-ramps)
◦ Traffic jam warning (speed reduction before traffic jam in sight)

Efficiency
◦ Congestion analysis & prediction
◦ Flow optimization & wave-breaking (AD & ADAS)
◦ Platooning (AD)

Traffic Patterns & Driving Behaviour
◦ Trajectory prediction / lane change / behavioural models for simulation
◦ Traffic flow and congestion predictions (performance single station vs. multiple stations)

Funding
Multiple funding options for the RABITS project are being considered:
◦ Direct assignment (by project partner)
◦ VINNOVA FFI
◦ Trafikverket Portfolio: Möjliggöra
◦ EU HORIZON-CL5-2024-D6-01-04 (as work package in larger project)

Partners & Project Roles

◦ RISE (Gothenburg, Stockholm): traffic state analysis, behavioural models, intent recognition, ML solutions
◦ KTH Traffic Lab (Stockholm): traffic forecasting, ML solutions
◦ VTI (Stockholm): ITS solutions, traffic models
◦ Trafikverket (Stockholm): data collection, data provision, problem owner
◦ NavTech Radar (UK): hardware & data owner, use case provider
◦ Volvo Cars (Gothenburg): participation sought, decision pending

6.5 Intention Recognition for multi-agent scenarios (WP 5)

Work Package 5	Intention Recognition for multi-agent scenarios
Responsible	HIS
Other participants	Volvo Cars, Smart Eye

1. Pedestrian intention recognition

Recognizing Pedestrians' intention to cross a street and predicting their imminent crossing action are major challenges for advanced driver assistance systems (ADAS) and Autonomous Vehicles (AV). During the project a master thesis project within data science was conducted on the topic of pedestrian intention recognition using feature fusion within a neural network architecture. In contrast to most of the earlier work on pedestrian intention recognition that focused either on using handcrafted features or on end-to-end deep learning approach, we investigate the impact of fusing handcrafted features with auto learned features by using a two-stream deep neural network architecture. The proposed method achieved very good results on the JAAD (Rasouli et al. 2017) dataset. Depending on if we considered only the immediate image frames before or image frames half a second before the crossing, we received prediction accuracy of 90%, and 84%, respectively (Hamed, 2020)

The work was extended and disseminated with two publications. In Hamed & Steinhauer (2021a) we addressed the pedestrian action prediction problem using the JAAD dataset and studied the impact of pedestrian-related and vehicle-related handcrafted features in a deep learning approach considering a scenario where we tried to predict the crossing action 0.5 seconds (15 frames) ahead of the actual action. In Hamed & Steinhauer (2021b) we substantially extended the approach by addressing both, intention recognition and action prediction. We propose a new neural network architecture and highlight the impact of context-related features and train and test our model on the large-scale dataset PIE (Rasouli et al. 2019) while also considerably extending the lookahead to 1.5 seconds (45 frames) ahead of the crossing action (or the car passing the pedestrian who is not crossing). This larger time frame is particularly needed when we want to predict a crossing intention rather than just the imminent crossing action. We evaluate our approach on the recently suggested benchmark by Rasouli et al. (2019) and show that our approach outperforms current state of the art models.

2. Learning individual driver's mental models

Advanced driver assistant systems are supposed to assist the driver and ensure their safety while at the same time providing a fulfilling driving experience that suits their individual driving styles. What a driver will do in any given traffic situation depends on the driver's mental model which describes how the driver perceives the observable aspects of the environment, interprets these aspects, and on the driver's goals and beliefs of applicable actions for the current situation. Understanding the driver's mental model has hence received great attention from researchers, where defining the driver's beliefs and goals is one of the greatest challenges. We attempted to establish individual drivers' temporal-spatial mental models of a situation by considering driving to be a continuous Partially Observable Markov Decision Process (POMDP) wherein the driver's mental model can be represented as a graph structure following the Bayesian Theory of Mind (BToM). The individual's mental model can then be automatically obtained through deep reinforcement learning. Using the driving simulator CARLA and deep Qlearning, we demonstrate our approach through the scenario of keeping the optimal time gap between the own vehicle and the vehicle in front (Darwish & Steinhauer, 2020).

3. State-of-the-art literature study of driver intention recognition

Dynamic Bayesian Networks, Hidden Markov Models and Deep Neural Networks have predominantly been applied to recognize driving maneuver intentions (Vellenga et al., 2022). Since 2016, DNNs have become the favorite approach for driver intention recognition, and the reliability and used dataset size increased. However, due to the costly and sensitive nature of collecting naturalistic driving data, most studies rely on proprietary datasets. This prohibits a general benchmark and comparison between the methods. There are multiple open-source datasets that can be used for driver intention recognition from multiple perspectives. For example, the Brain4Cars dataset consists of more than 1000 miles of in-cabin and external camera footage collected in San Franscisco (Jain et al., 2016), and the Honda Research Institute Driving Dataset (HDD, Ramanishka et al., 2018) consist of 104 hours of lane branch, lane change, merge, park, passing, and turn maneuvers and includes camera, vehicle dynamics and LiDAR observations.

4. Understanding and Estimating Uncertainties for Driver Intention Recognition

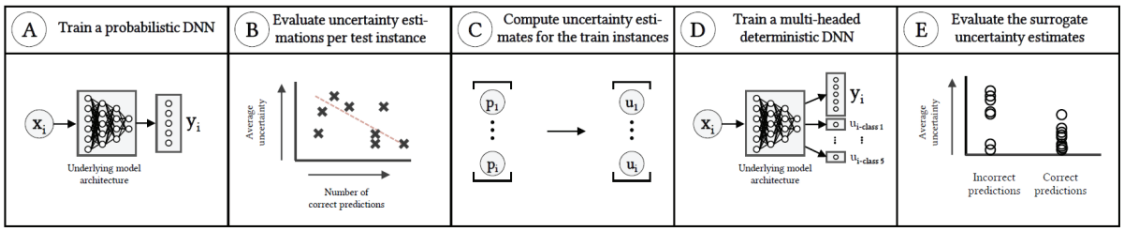

Figure 13. Schematic overview of the surrogate uncertainty estimation approach

Applying artificial intelligence (AI) in a safety-critical context like driving requires a careful approach. One of the current challenges of AI is to handle unseen and rare scenarios in a safe way (Hendrycks et al. 2021). Moreover, regular deep neural networks (DNNs) do not produce uncertainty estimations about produced predictions. In line with upcoming AI regulations (e.g., EU, USA, Japan), a driver intention recognition (DIR) system should be able to quantify how certain it is about the produced predictions. Uncertainties originate from several sources, such as measurement errors, statistical model induction, regularization effects (Darling, 2019). To yield uncertainty estimates from a regular DNN, one can use probabilistic DNNs (PDNNs, Gawlikowski et al. 2021). A downside of estimating the uncertainty for every instance while making use of a PDNN is that multiple inferences for a single instance are required. To overcome this computationally expensive property of PDNNs, we tested for multiple probabilistic deep learning (PDL) methods (Variational Inference, Monte-Carlo Dropout, Stochastic Weight Averaging – Gaussian and a Deep Ensemble) to understand if we can learn to reproduce the uncertainty estimations with a surrogate model. For each of the PDL methods we estimate the uncertainty for the train instances and train a multi-headed deterministic DNN to simultaneously recognize the driving intentions and estimate the uncertainty. In Vellenga et. al (2023) we found that the surrogate models based on the Monte-Carlo dropout and Variational Inference models produced significantly higher uncertainty estimates for incorrectly classified test instances. In practice this would mean that, if test properly, one would not have to perform multiple inferences for each instance and still get an uncertainty estimate.

5. Evaluation of deep neural network design for driver intention recognition.

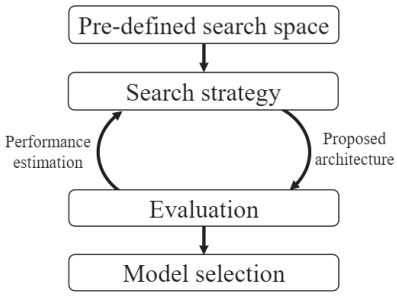

Figure 14. Overview of the Neural Architecture Framework

Given the limited computational capabilities available in a car and that the majority of most recent DIR studies rely on DNNs to achieve state-of-the-art performance (Xing et al. 2020, Rong et. al 2021, Ma et al. 2023), we investigated how to motivate the design of the network most effectively. We performed a neural architecture search (NAS) for three DNN layer types (an LSTM, TCN or TST), using different information fusion strategies to combine data from different sensors, and eight search optimization strategies (Random Search, Hill-Climbing, Particle Swarm Optimization, Regularized Evolution, Gaussian Process, Tree-structured Parzen Estimators, Policy based reinforcement learning and Latent Action Neural Architecture search) on two DIR datasets. NAS consists of defining a search space, which we build by considering the depth and width of the networks. Afterwards, the search strategy iteratively samples networks from the search space and aims to achieve the best value for a set objective, such as accuracy or performance. Compared to the manually designed models, we observed improved performance for the NAS sampled architectures. Moreover, for both datasets we found that a more complex model does not necessarily lead to improved performance.

6. State-of-the-art video-based driver action and intention recognition

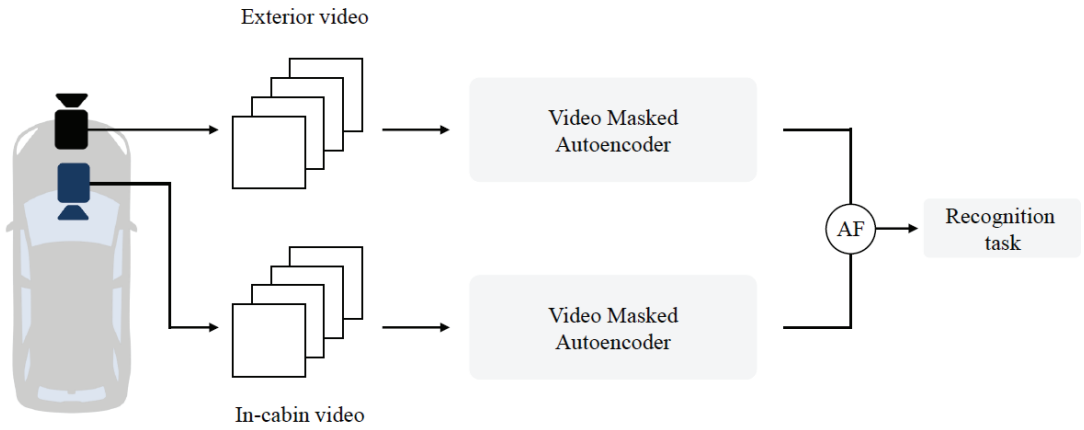

Figure 15. Overview of the multi-video setup. Video representations are learned by two video masked auto encoders, after which attention fusion (AF) is performed to create a joint embedding.

Traffic fatalities remain among the leading death causes worldwide. To improve road safety, car safety is listed as one of the most important factors. To actively support human drivers, it is essential for advanced driving assistance systems to be able to recognize the driver's actions and intentions.

Prior studies have demonstrated various approaches to recognize driving actions and intentions based on in-cabin and external video footage. Given the performance of self-supervised video pre-trained (SSVP) Video Masked Autoencoders (VMAEs) on multiple action recognition datasets, we evaluate the performance of SSVP VMAEs on the Honda Research Institute Driving Dataset for driver action recognition (DAR) and on the Brain4Cars dataset for driver intention recognition (DIR). Besides the performance, the application of an artificial intelligence system in a safety-critical environment must be capable to express when it is uncertain about the produced results. Therefore, we also analyzed uncertainty estimations produced by a Bayes-by-Backprop last-layer (BBB-LL) and Monte-Carlo (MC) dropout variants of an VMAE. Our experiments show that an VMAE achieves a higher overall performance for both offline DAR and end-to-end DIR compared to the state-of-the-art. The analysis of the BBB-LL and MC dropout models show higher uncertainty estimates for incorrectly classified test instances compared to correctly predicted test instances.

7. **An intention recognition framework for modelling, inferring and predicting intentions for the chosen use cases based on scenario.**

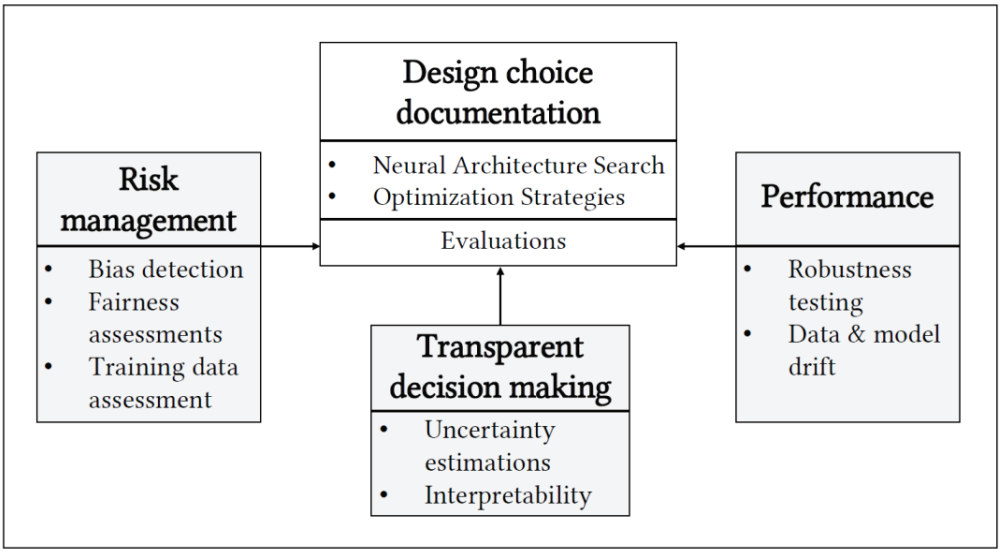

Figure 16. Foundational framework centered around the requirement to provide design choice documentation for a high-risk AI-based DIR system.

Previous studies have proposed frameworks to guide the safe design, validation and evaluation of AI-based systems (e.g., Salay and Czarnecki, 2019; Pereira and Thomas, 2020; Mock et al., 2021; Häring et al., 2021). While these frameworks provide essential steps for safe AI system development, Tarrisse and Massé (2021) and Neto et al. (2022) argue that most of these initiatives are still work-in-progress and lack details about methods and about the influence of the processes on each other. Therefore, and drawing from the insights of the included papers, we formalize and complement previous efforts to guide safe high-risk AI system design and integration in Figure 12. This framework centers around the upcoming requirement to motivate the design choice of a high-risk AI-based system and how it affects the performance. Given the evolving nature of technology and ML methodologies, it is necessary to periodically review and extend the framework. Initially, we include three dimensions that should be considered when empirically motivating the design. The processes of this framework consist of system design motivation, transparent decision-making, risk management, and performance monitoring.

6.6 Exploitation and Business Value (WP 6)

Work Package 6	Exploitation and Business Value
Responsible	Volvo Cars
Other participants	SEYE, TrV

Possible applications
The domain of intention recognition has a wide range of applications in the automotive industry. The following is a brief description of few possible applications of intention recognition systems in the automotive industry.

- Advanced Driver Assistance Systems (ADAS): Intention recognition can be used to predict the driver's intentions behind certain actions. By analysing their behaviour, such as gaze direction, hand movements, and body posture, systems can better anticipate manoeuvres like lane changes, turns, or braking, thereby improving the safety measures and actions taken by of the ADAS systems.

- Driver Monitoring Systems: Intention recognition methods can be integrated into technologies where driver monitoring systems are used to detect driver distraction, fatigue, or drowsiness. By analysing the eye movements, facial expressions, and body posture, it could be possible to infer the driver's intentions or mental state. These methods in combination with other sensor fusion technologies could be used for triggering alerts that could improve driver attention or take interventions when necessary.

- Personalized Driving Experience: Intention recognition can help in creating a more personalized driving experience. For example, the intentions of the drivers or the occupants to use certain comfort and convenience functions could be developed based on historical patterns in usage of the functions. These in combination with external data such as the preferences of the driver or occupants in certain situations could further fine tune the personalization predictions. Furthermore, the intention recognition methods in personalisation could be used to learn and adapt to a driver's habits, preferences, and intentions including but not limited to adjusting settings such as seat position, climate control, or infotainment options etc.

Industrial PhD as bridge between academia and industry
The Volvo Industrial PhD program (VIPP) serve as a bridge between academia and industry, fostering collaboration and knowledge exchange between Volvo cars and the academia. The VIPP program typically involves a doctoral student conducting research within the industrial environment while being supervised by both academic and industry mentors. The VIPP program emphasizes both methodological and applied research areas that addresses real-world problems faced by the automotive industry in general and also specific to the ambitions of Volvo cars. By working closely with both academia and industry, the industrial PhD candidates facilitate the transfer of knowledge, methodologies, and best practices between these two domains in both directions. The goal is that the industrial PhD candidates bring academic expertise into the industrial setting and identifying relevant research methodologies applicable to the practical challenges faced by Volvo cars. In this context of intention recognition, the project has provided the industrial PhD student access to resources, facilities, infrastructure, and data that is relevant for

developing methods for intention recognition that are relevant to use cases in the automotive industry. These enabled the collaboration between the academia and industry to conduct research in context of automotive industry relevance. This nature of research is one step closer towards enabling industrialisation of developed methods which have led to more applicable outcomes. The current industrial research had focus on practical applications of intention recognition in the automotive industry and collaboration between academia and industry.

Production readiness and challenges

Although intention recognition methods have significant potential applications, the industrialization of intention recognition methods in the automotive industry presents several challenges. The methods need to accurately interpret various signals and cues from drivers, pedestrians, and other vehicles which remains a challenge especially in complex real-world traffic and user contexts. High quality and reliable datasets are key enablers for the intention recognition models. Such diverse datasets constitute a wide range of scenarios, behaviours and environmental conditions. Identifying and collecting diverse datasets is both challenging and time consuming. The intention recognition models, and the systems must be robust and should be able to handle uncertainties and edge cases. Furthermore, the functions employing the intention recognition models shall handle and adapt to variations in driving behaviour and environmental changes. When it comes to personalisation applications, the systems may also need to adapt to driving styles, cultural aspects of driving. Collection and processing data related to human intentions may raise concerns about privacy and ethical aspects. Therefore, the industry needs to balance the need for data collection while ensuring privacy of the end users and thus gaining trust in the data collection environments is a significant challenge. Deploying the intention recognition models or systems with intention recognition methods into vehicles involves complying with the existing safety and regulatory standards within the automotive industry. Thus, the safety and reliability aspects of system employing intention recognition methods or frameworks, and their integration into existing vehicle safety functions are very critical.

While complete intention recognition algorithms are not yet ready for production programs, during this project, Smart Eye has made big progress in driver and cabin monitoring signals. That not only can support intention recognition algorithms in the future but is something that will be moved into production development to modify and complement output signals of Smart Eye products.

6.7 Demonstration and Proof-of-Concept (WP 7)

Work Package 7	Demonstration and Proof-of-Concept
Responsible	Volvo Cars
Other participants	SEYE, TrV, RISE

Below a brief overview of the various partner seminars, workshops, and demos as part of the IRRA-project.

Topic	**Presenter**
IRRA definition of intention	H. Joe Steinhauer
Methods for intention recognition	Koen Vellenga
Progress in the logic-based intention recognition approach	Björn Bjurling
Learning individual drivers' mental models using POMDs and BToM	Amena Darwish
IRRA use case workshops	All
Smart Eyes´ Interior Sensing Systems: Driver and interior monitoring. Data collection capabilities and potential insights	Svitlana Finer

7. Dissemination and publications

7.1 Dissemination

How are the project results planned to be used and disseminated?	Mark with X	Comment
Increase knowledge in the field	X	
Be passed on to other advanced technological development projects	X	
Be passed on to product development projects	X	
Introduced on the market		Remains difficult given the technology state and safety requirements from both the industry and regulators.
Used in investigations / regulatory / licensing / political decisions	X	

7.2 Publications

- Darwish, A., & Steinhauer, H. J. (2020). Learning individual driver's mental models using POMDPs and BToM. In Proceedings of the 6th International Digital Human Modeling Symposium (pp. 51-60).
- Hamed, O. (2020) Pedestrian Intention Recognition: Fusion of Handcrafted Features in a Deep Learning Approach. Independent thesis Advanced level (degree of Master (One Year)), 10 credits / 15 HE credits, University of Skövde
- Hamed, O., & Steinhauer, H. J. (2021). Pedestrian's Intention Recognition, Fusion of Handcrafted Features in a Deep Learning Approach. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 18, pp. 15795-15796).
- Hamed, O. & Steinhauer, H. J. (2021b). Pedestrian Intention Recognition and Action Prediction Using a Feature Fusion Deep Learning Approach. The 18th International Conference on Modeling Decisions for Artificial Intelligence: MDAI
- Vellenga, K., Steinhauer, H. J., Karlsson, A., Falkman, G., Rhodin, A., & Koppisetty, A. C. (2022). Driver intention recognition: State-of-the-art review. IEEE Open Journal of Intelligent Transportation Systems.
- Vellenga, K., Karlsson, A., Steinhauer, H. J., Falkman, G., & Sjogren, A. (2023). Surrogate Deep Learning to Estimate Uncertainties for Driver Intention Recognition. In Proceedings of the 2023 15th International Conference on Machine Learning and Computing (pp. 252-258).
- Vellenga, K., Steinhauer, H. J., Falkman, G., Björklund, T. (2024). Evaluation of Video Masked Autoencoders' Performance and Uncertainty Estimations for Driver Action and Intention Recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 7429-7437).
- Vellenga, K., Steinhauer, H. J., Karlsson, A., Falkman, G., Rhodin, A., & Koppisetty, A. C. (2024). Designing deep neural networks for driver intention recognition. Under submission to: Engineering applications of artificial intelligence (Elsevier journal).
- Vellenga, K. (2024). Advancing Deep Learning-based Driver Intention Recognition: Towards a safe integration framework of high-risk AI systems. Licentiate Thesis, University of Skövde.

8. Conclusions and future research

We conclude our work in the IRRA project by reflecting upon the work across the various work packages (WPs). Each WP targeted specific aspects of intention recognition, ranging from data collection to advanced algorithm development, and testing their integration for practical applications. In this section, we summarize the key findings and advancements made in each WP, highlighting how they collectively contribute to advancing the field of intention recognition in the automotive sector.

WP 1: Data Collection and Procurement
Key Achievements: WP 1 laid the groundwork by establishing robust data collection methodologies. Despite challenges in data transfer and synchronization, effective strategies were developed for gathering and processing real-time vehicular and environmental data, essential for enhancing intention recognition algorithms and algorithms supporting them. Additionally, to the obtaining data through data collection, other ways of data procurement were established to ensure that needs of the project are met.

Implications: This foundational work ensures that subsequent packages are fed with high-quality, relevant data, crucial for the development of accurate and efficient intention recognition systems.

WP 2: Driver and Environment Monitoring
Achievements: WP 2 delivered a framework for extracting necessary features from images. This involves combining existing pre-trained methods and fine-tuning for a given task. Additionally, we have explored the potential of training new models to extract features that are not typically captured, such as brake light detection.

Implications: The analysis based on pre-trained networks achieved reasonable results and the extracted features can be utilized in WP3. However, the extraction of certain non-typical features (e.g., brake lights) requires adequately labelled data. The collection or generation of such labelled images is therefore a logical next step. Moreover, the length of the historical context should be extended for more accurate and robust models.

WP 3: Vehicle Perspective: Intentions of Drivers of Interacting Vehicles
Achievements: WP3 introduces logical and statistical models designed to understand the current active situation and to reason about upcoming potential scenarios.

Implications: A more diverse set of detected features would enable improved generalizability of models such as those in WP2 across various scenarios. Additionally, accuracy and robustness of the models could be enhanced by integrating additional data sources such as map data, which provide context beyond the features visible in camera images.

WP 4: System Perspective of Intentions
Achievements: Due to the restructuring of this WP the work was confined to the preparation of a follow-up project. To this end, a comprehensive definition of system-level intention recognition was developed, a consortium was built, access to a large and relevant dataset secured, and funding opportunities were explored.

Implications: The consortium will finalize a project application to one of the named tenders. By introducing the idea of collaborative or shared perception to intention recognition, the sought RABITS project has great potential to enhance the robustness and effectiveness of intention recognition in AD and ADAS systems.

WP 5: Intention Recognition for Multi-Agent Scenarios
Achievements: This WP tackled multiple complex intention recognition scenarios involving multiple agents, such as pedestrian intention recognition and ego-vehicle driver intention recognition as well as individual driver mental model analysis. Novel deep neural network architectures and reinforcement learning models achieved state-of-the-art performance on multiple intention recognition benchmarks, which represent the contribution of the IRRA project to significant advancements in the field.

Implications: The outcomes of this WP do not only advance the technical capabilities in intention recognition but also set a precedent for integrating regulatory frameworks within AI development in the automotive sector. This holistic approach ensures that the technological advancements are robust, ethically sound, and compliant with regulatory standards, paving the way for safer and more efficient implementation of AI in critical domains.

WP 6: Exploitation and Business Value
Achievements: This work package performed a feasibility assessment, affirming the applicability of intention recognition research in the automotive industry. It highlights the industrial relevance of the research, identified mechanisms for effective knowledge transfer, and evaluated the production readiness of intention recognition systems, ensuring their practical integration into the automotive sector.

Implications: The implications of this work are profound, particularly in enhancing safety measures and reshaping the driving experience. The successful application of intention recognition in, for example, Advanced Driver Assistance Systems (ADAS) promises proactive responses to driver manoeuvres, improving overall safety. Additionally, integration with Driver Monitoring Systems and personalized driving experiences underscore the potential for transformative advancements, setting new standards in the automotive industry and influencing future technological developments.

Overall Conclusion
The IRRA project's approach to intention recognition encompassed a wide range of methodologies and applications and has significantly advanced the understanding and capabilities in this domain. Each WP contributed to this goal, from foundational data collection to sophisticated modeling of intentions in complex scenarios. These results not only enhance the field of intention recognition but also have practical implications, particularly in improving safety and efficiency in the automotive sector.

9. Participating parties and contact persons.

Academic partners:
HIS – H. Joe Steinhauer
HIS – Göran Falkman
HIS – Alexander Karlsson

RISE – Leon Sütfeld (leon.suetfeld@ri.se)
RISE – Björn Bjurling (bjorn.bjurling@ri.se)
RISE – Anders Holst (anders.holst@ri.se)
RISE – Sepideh Pashami (sepideh.pashami@ri.se)
RISE – Lars Rasmusson (lars.rasmusson@ri.se)
RISE – Åsa Rudström (no longer at RISE)
RISE – Henrik Malmgren (no longer at RISE)

Industry:
Smart Eye AB – Henrik Lind (henrik.lind@smarteye.se)
Smart Eye AB – Svitlana Finér (svtilana.finer@smarteye.se)
Smart Eye AB – Raimondas Zemblys (raimondas.zemblys@smarteye.se)

Volvo Car Corporation – Koen Vellenga
Volvo Car Corporation – Ashok Koppisetty
Volvo Car Corporation – Hampus Grimmemyhr
Volvo Car Corporation – Asli Rhodin (no longer at VCC)
Volvo Car Corporation – Ivana Jern (no longer at VCC)
Volvo Car Corporation – Koshan Emani (no longer at VCC)

Industrial partner:
Trafikverket – Tomas Julner (tomas.julner@trafikverket.se)

References:

- EU Commission et al. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.
- Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., ... & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, *56*(Suppl 1), 1513-1589.
- Gillblad, D., Steinert, R. and Holst, A. (2008). Fault-tolerant incremental diagnosis with limited historical data. In International Conference on Prognostics and Health Management 2008 (PHM'08).
- Hendrycks, D., Carlini, N., Schulman, J. and Steinhardt, J. (2021). Unsolved problems in ML safety. arXiv preprint arXiv:2109.13916
- Jain, A., Koppula, H. S., Soh, S., Raghavan, B., Singh, A., & Saxena, A. (2016). Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture. *arXiv preprint arXiv:1601.00740*.
- Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., ... & Zhang, L. (2023). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499.
- Ma, Y., Ye, W., Cao, X., Abdelraouf, A., Han, K., Gupta, R., & Wang, Z. (2023). CEMFormer: Learning to Predict Driver Intentions from In-Cabin and External Cameras via Spatial-Temporal Transformers. *arXiv preprint arXiv:2305.07840*.
- Michael C Darling. (2019). Using Uncertainty To Interpret Supervised Machine Learning Predictions.
- Ramanishka, V., Chen, Y. T., Misu, T., & Saenko, K. (2018). Toward driving scene understanding: A dataset for learning driver behaviour and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7699-7707).
- Rong, Y., Akata, Z., & Kasneci, E. (2020). Driver intention anticipation based on in-cabin and driving scene monitoring. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1-8). IEEE.
- Sadri, Fariba. "Logic-based approaches to intention recognition." Handbook of research on ambient intelligence and smart environments: Trends and perspectives. IGI Global, 2011. 346-375.
- Sergot, Marek. "A computational theory of normative positions." ACM Transactions on Computational Logic (TOCL) 2.4 (2001): 581-622.
- Vellenga, K., Karlsson, A., Steinhauer, H. J., Falkman, G., & Sjogren, A. (2023). Surrogate Deep Learning to Estimate Uncertainties for Driver Intention Recognition. In *Proceedings of the 2023 15th International Conference on Machine Learning and Computing* (pp. 252-258).
- Vellenga, K., Steinhauer, H. J., Karlsson, A., Falkman, G., Rhodin, A., & Koppisetty, A. C. (2022). Driver intention recognition: State-of-the-art review. *IEEE Open Journal of Intelligent Transportation Systems*.
- Xing, Y., Lv, C., Wang, H., Cao, D., & Velenis, E. (2020). An ensemble deep learning approach for driver lane change intention inference. *Transportation Research Part C: Emerging Technologies*, *115*, 102615.
- Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., ... & Wang, X. (2022, October). Bytetrack: Multi-object tracking by associating every detection box. In European Conference on Computer Vision (pp. 1-21). Cham: Springer Nature Switzerland.