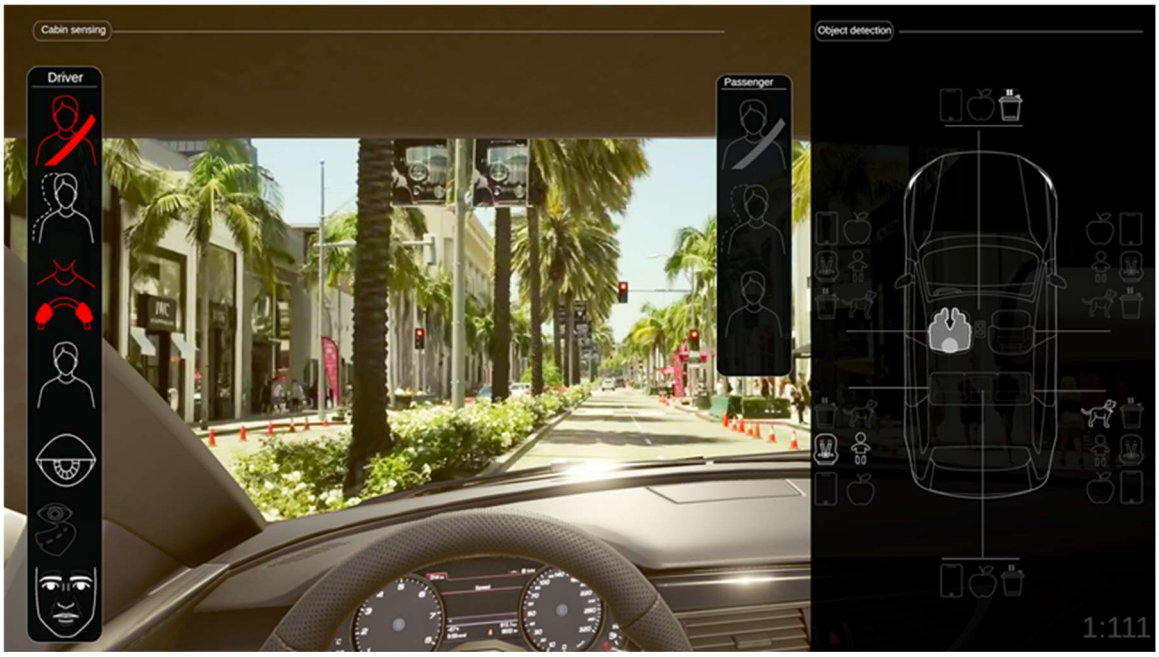
DRAMA

Public report



Project within EMK
Author Thanh Hai Bui, Martin Torstensson, Henrik Lind, Svitlana Finér
Date 20200227

FFI Fordonsstrategisk Forskning och Innovation

Content

1 Summary

The DRAMA project develops knowledge within driver activity mapping to improve interaction between semi-/automated vehicles and human drivers/passengers - with the overall goal to improve traffic safety and driver/passenger comfort. 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 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. The decision if the vehicle can perform a handover or a safe stop can be determined based on in-vehicle activity and driver attention/disengagement in the driving task.

Recognition of human behaviour within vehicles are becoming increasingly important. Paradoxically, the more control the car has (i.e. in terms of ADAS), the more we need to know about the person behind the wheel [1] especially if he/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 comfort 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 reengage driving task at some point, it will be important to know more about what happens inside the vehicle. In this report we describe the work performed within the DRAMA project to find methods/algorithms/tools that can be used to map driver/passenger body posture/facial expressions/object/etc. and combine them in activity/behaviour models. The project also performs research and implements proof of concept prototypes for a camera-based data acquisition system to be able to capture the activity/behaviour in realtime.

2 Sammanfattning på svenska

DRAMA-projektet utvecklar kunskap inom kartläggning av föraraktiviteter för att förbättra interaktionen mellan helt eller halvt automatiserade fordon och mänskliga förare eller passagerare - med det övergripande målet att förbättra trafiksäkerheten och förarens / passagerarnas komfort. Med korrekta klassificeringar av vad föraren gör kan interaktion mellan människa och maskin riktas till den mest lämpliga modaliteten (visuellt / ljud / haptiskt) vid en given tidpunkt. Om bilen känner till passagerarnas kroppspositioner, kan säkerhetsfunktionerna anpassas för att användas optimalt i sammanhanget, till exempel airbags, styrning, broms och undvikande manöver. Beslutet om fordonet kan göra en

överlämning eller ska utföra ett säkert stopp kan bestämmas utifrån aktiviteter i fordonet och förarensuppmärksamhet / oengagemang i köruppgiften.

Igenkännande av mänskligt beteende inom fordon blir allt viktigare. Paradoxalt nog, ju mer kontroll bilen har (dvs när det gäller stödsystem), desto mer behöver vi veta om personen bakom ratten [1], särskilt om han eller hon förväntas ta kontroll över automatiseringen. Mycket fokus har ägnats åt forskning om sensorer som övervakar den yttre miljön, men sensorer på insidan har inte fått i närheten av lika mycket uppmärksamhet. När det gäller övervakning av distraktioner kan vad som för närvarande ses som farligt (t.ex. användning av mobiltelefoner) i framtiden ses som något bra som hjälper till att hålla människor vakna i mycket automatiserade fordon. En annan anledning till att kartlägga aktiviteter inuti bilen är skillnaden mellan förarens förväntningar och verkligheten i vad dagens automatiska fordon kan [2]. Så länge automatiseringen kommer med begränsningar som kräver att föraren tar över kontroll vid något tillfälle, är det viktigt att veta mer om vad som händer i fordonet. I denna rapport beskriver vi det arbete som utförts inom DRAMA-projektet för att hitta metoder/algoritmer/verktyg som kan användas för att kartlägga förare/passagerarkroppsställning/ansiktsuttryck/objekt etc. och kombinera dem i aktivitets-/beteendemodeller. Projektet utför också forskning och implementerar ett koncept för ett kamerabaserat datainsamlingssystem för att kunna fånga upp aktiviteter/ beteenden.

3 Background

An enabler for driver and passenger monitoring for improved user experience and safety is camera-based interior sensing. A camera solution is more flexible than for example pressure sensors in the seats to measure sitting position or wearable glasses for measuring gaze. Camera based monitoring is non-intrusive in terms of relieving test persons of having devices and sensors attached to their body. Recent findings even show that an off-the-shelf camera can measure the pulse of a person remotely [3]. Thus, there is large potential in using such systems for mapping of driver and passenger activity in relation to both UX and safety functions.

The purpose of the DRAMA project is to enable accurate mapping and modelling of human in-vehicle behaviour using computer vision and machine learning to enable improved safety and UX functionality.

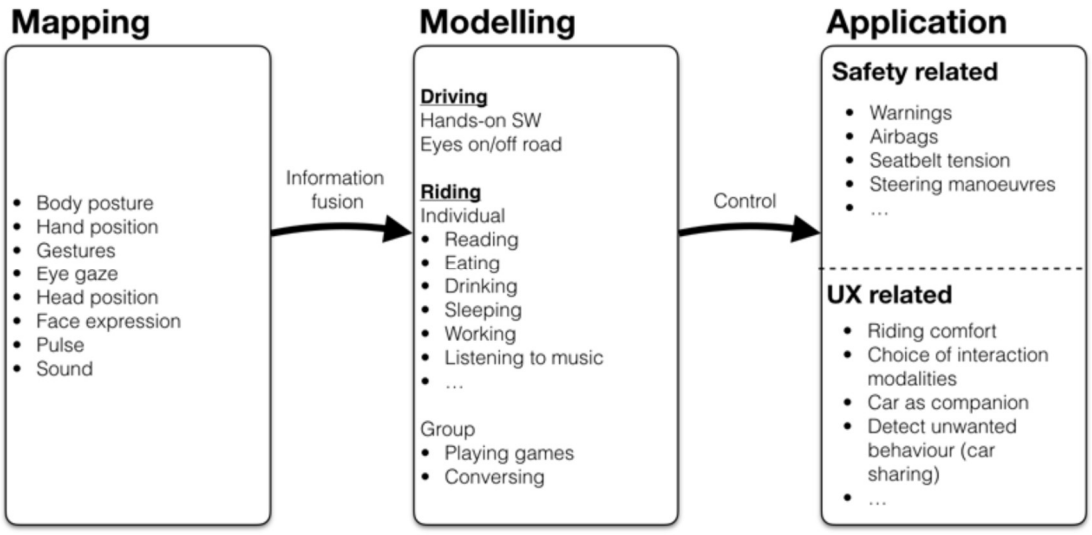


Figure 1: Mapping and modelling of activities is an enabler of new functions in the vehicle. In the DRAMA project the main focus is on the initial mapping and modelling

The project contributes to more efficient development and evaluation of applications related to automated driving where the human has a role. Consequently, it strengthens the offer by enriching product portfolio from Smart Eye and strengthen its international competitiveness. In addition, it helps RISE to develop expertise that is applicable in current and future projects, that will gain RISE mission to support the development of Swedish industry. This project develops software necessary for realising mapping and modelling of the activities within the vehicle cabin based on computer vision using cameras.

4 Purpose, research questions and method

Driver behaviour is the most important factor for road traffic safety [4]–[6]. The higher automation level the car has, the more understandings between the car and the people inside the cabin will be required to create a safe and comfort journey, especially when handover of driving modes is still needed.

The DRAMA project provides knowledge within three main areas that are tightly connected.

a) Mapping algorithms that detect emotions, postures, head and eye tracking, faces and objects.
b) Architectures for video data acquisition and processing
c) Modelling algorithms and software, based on machine learning, to create the models that transform visual features from different recognition components into driver/passenger activities/interactions and states.

The following research questions are investigated within the project:

- RQ1: What are the in-cabin activity scenarios related to safety and comfort of a car journey?
- RQ2: How one can derive mapping and modelling of in cabin activities from camera data to improve safety/comfort?
- RQ3: How can we setup data acquisition system for initial datasets and for in-cabin system?

The following methodology is applied:

- Study the industry requirements, literature and surveys of in-cabin activities that are of interested and related to safety and comfort in SAE2-5 highly automated vehicles.
- Generalize and derive in-cabin activity scenarios structure so that the system is scalable and capable to accommodate new/modified scenarios with constant or linear time complexity.
- Design interior sensing algorithms using deep learning and computer vision techniques to (i) extract features representing different aspects of in-cabin activities and (ii) classification methods that can classify input image sequences into pre-defined classes of in-cabin activities/interactions.
- Capture training data under different environment settings to evaluate and figure out techniques to increase the robustness of the trained networks against variations of environment (illumination conditions, camera perspectives, etc)
- Prototyping and evaluating the developed concept.

5 Objective

The DRAMA project has the following objectives supporting overall FFI objectives

Increasing the Swedish capacity for research and innovation, thereby ensuring competitiveness and jobs in the field of vehicle industry

The DRAMA project contributes to Swedish industry with state-of-the-art driver monitoring and activity modelling tools to be able to develop automated vehicles that are capable of interacting with the driver and passengers.

Developing internationally interconnected and competitive research and innovation environments in Sweden

Smart Eye has partnered with Ambarella[7] to deliver next generation AI-based Driver Monitoring. Smart Eye also integrates the algorithms developed during DRAMA project into its next generation product line of interior sensing.

DRAMA works have also been presented at the following conferences around the world:

- The 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM), 2019. Prague, Czech Republic
- The Third Swedish Symposium on Deep Learning (ISDL), 2019. Norrköping, Sweden
- 14th IEEE International Conference on Automatic Face Gesture Recognition (FG), 2019. Lille, France
- CES 2020, Las Vegas, NV.
- SAFER pre-event at 3rd Global Ministerial Conference on Road Safety (UN), 2020. Stockholm, Sweden.

Promoting the participation of small and medium-sized companies

Smart Eye is an SME providing innovative products and services to the car OEMs and Tier 1s. The project helps Smart Eye to enrich the product portfolio and initiate its interior sensing product line.

Promoting cross-industrial cooperation
DRAMA project attracts interests of other SMEs who are also working in deep learning solutions for different industries, e.g. Berge Consulting AB who will promote the DRAMA results to customers from other industries. We also discuss collaboration with Assistant Prof. Eren Aksoy (Halmstad University) similar applications in robotics, where robots can learn activity cues from human to perform assigned complex tasks.

Promoting cooperation between industry, universities and higher education institutions

The project helps to deliver 2 diploma theses lead by RISE: Martin Torstensson, Erik Kratz, and a summer student job Carmen Lee. As well as 1 Master's thesis lead by Smart Eye: "A Deep Learning based tracking framework for passenger monitoring" in Computer Science – algorithms, languages & logic and Computer Systems & Networks, carried out by students Filip Granqvist and Oskar Holmberg in 2018.

Promoting cooperation between different OEMs

The DRAMA project is a method and software project. In the next project when the methods are applied to gain knowledge about the driver behaviour there will be a natural environment for OEM collaboration.

Knowledge and competence development at research institutes and companies

The project contributes to increasing the competence within software, algorithms and machine learning in the area of mapping human activity/behaviour.

Personalized functionality

By detecting the emotional state of the driver, different interaction strategies may be applied to communicate with the driver. By knowing the pose/sitting position of the driver/passengers, the safety measures may be different in case of an emergency. The framework developed within DRAMA will thus enable personalized functionalities for prospect safety and comfort systems.

Smart functions and intelligent assistance in vehicle specific domains

The development of a modular multi-core concept system for data collection and real-time data analysis in the vehicle contributes to this focus area. Data collection and data analysis system can provide information to other agents active in the driving situation, both agents/functions in the own vehicle but also within vehicles in the surrounding.
To this end, the project also coordinates with on-going national initiatives including: TIC (FFI 2017-03058), HARMONISE (FFI, 2016-04263) and SMILE II (FFI, 2017-03066).
Being visible at international conferences is also prioritized, we have the project results disseminated in conferences in Sweden, Czech, France and the US (see Section 7.1).

The project creates an innovation foundation for further innovation projects in the area, typical examples including estimation of real-time comfort level, distraction level, re-engagement time (for handing over the driving task).

6 Results and deliverables

6.1 WP1 - Scenario description and selection and prioritization

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 gestures and body position, the 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.

In order to provide a set of scenarios for further study in WP3/4, we use the following approach:

- Gathering business requirements for interior sensing from industry. The preliminary list is provided in Appendix 10.1.1
- Literature review: We conduct in-depth 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 (for vehicles of SAE 2-3)[8]–[11] and (ii) literature on most common activities in autonomous vehicles (SAE 4-5)[12]–[16]. Since AV today mainly fall into SAE2-3 categories, there is no report for SAE 4-5 AV exist based on statistics, but from the survey of what people want to do in autonomous cars. The scenarios are thus categorized into comfort- and safety-related lists (Appendix 10.1.2)
- A finalized list of scenarios has then been reviewed and categorized through 3 design workshops involving multidisciplinary researchers, with regards also to the availability of related recognition algorithms, applying the following design principles:
    - It is a minimal spanning set where each atomic component is representing by at least 2 instances to avoid unintended dependency of the recognition on any component instance.
    - It spans all potential scenarios including the compiled lists (by previous steps) by generalizing the instances, not component. This is to ensure that the system is scalable and can address existing as well as foreseen future requirements for interior sensing. Technically speaking, it is equivalent to the hypothesis that newly arrival scenarios (with the same input settings) can be accommodated by retraining the related part of the system will new annotated data without the need to modify the algorithm architecture or complete retraining.

The final list of scenarios (step 3) is provided in Appendix 10.3, consisting of 56 detailed scenarios where permutations are performed with all three dimensions (variables): actors, activities and objects. The 56 scenarios are created with the following component instances (representative domain values of the independent variables):

- Seat position: Driver, front-seat passenger, back-seat passenger (1-2).
- Object: Apple, banana, phone, bottle, cup, laptop, steering wheel.
- Activity/interaction: Conversation, handshake, high-five, handling, showing, eating, drinking, holding, talking, using, reading, reaching and grabbing, arbitrary other activity.

Each component instance will appear in scenario list together with at least 2 instances of other components. This setup will remove unintentional correlation between a scenario

and any specific instance, e.g. if “orange” object will result in the prediction of “driver is handing over an orange to front seat passenger”.

We select component architecture using triple statement describing the activity (Subject, Verb predicate, Object), spanning the space of basic context independent activities/interactions. Future researches may also investigate adding more complex interior and exterior traffic contexts, using the same approach.

In parallel with the literature review on activity scenarios, we also conduct another literature review on the state-of-the-art algorithms, public datasets, and pre-trained weights. This review provides valuable input for the architecture design in technical feasibility perspective. This helps to ensure that the architecture enjoys state of the art algorithm developments for its components.

State of the art algorithms within HAR are falling into one or several categories as listed below (details are tabulated in Appendix 10.2):

- Hierarchical model
- Multiview
- Handcrafted feature based
- Deep learning
- Combine: fusion in different phase in the network

Challenges: No publicly available datasets for activity recognition, available activity recognition algorithms do not cover the project interested activity list and work in different environment (datasets).

6.2 WP2 - Modular multi core concept system

The data acquisition system is developed based on the knowhow and technologies proven by Smart Eye data capture system.

The system consists of a processing unit connected with multiple input visual sensors via

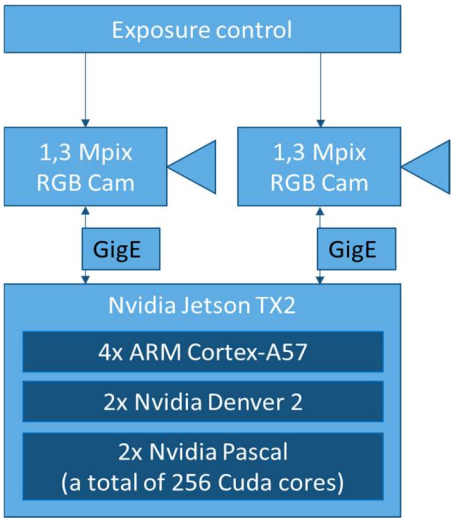


Figure 2: Data acquisition architecture

GigE Vision interface standard (Figure 2). The mounting setup of input sensors are provided in Figure 3, providing FoV covering all car seats (of a typical 5 seat passenger car). In Figure 3 one can see one position for the front camera and two positions for the back camera. During this project both positions were considered and tested.
Selected input visual sensors (Figure 4): Multiple IR enabled RGB cameras (Setup 1: 1.3M pixel with 150º FoV lens, Setup 2: 2M pixel RGB-IR cameras with 185º FoV lens) installed at the front and passenger seat coverage.

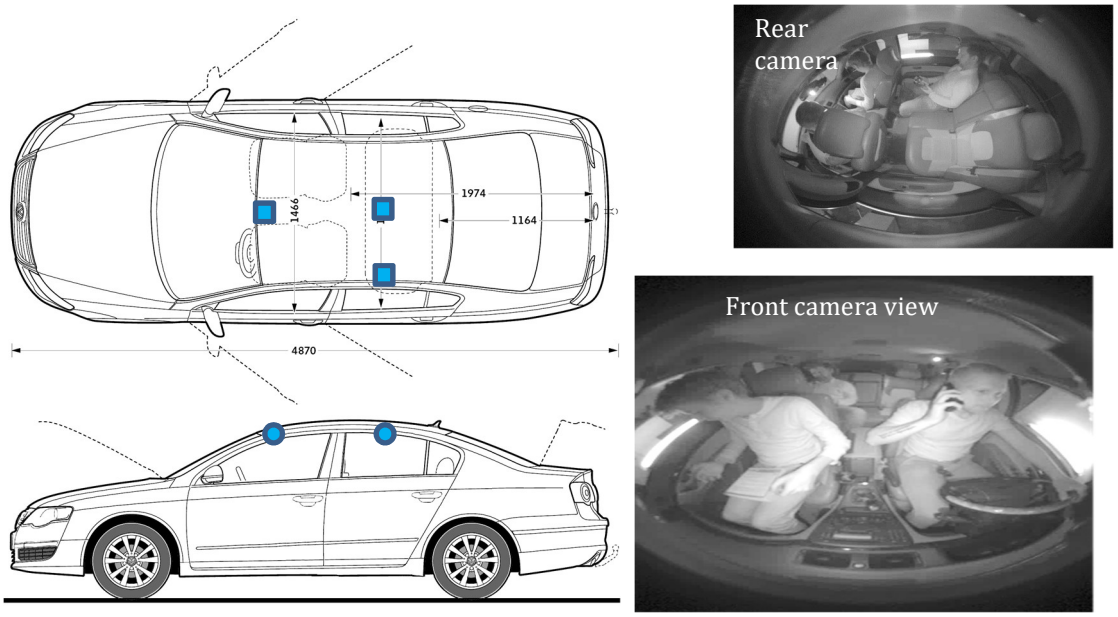


Figure 3: Camera mounting positions in car cabin

The motivation of selecting RGB-IR cameras is based on the followings:

- It is possible to choose whether to run algorithms using RGB image or IR image or both.
- It does not depend on ambient light, providing output independently whether it is night or day, tunnel or open road.
- It can also be used for video conference calls.
- Combining two modalities will provide more accuracy, because some materials are not very visible in IR so one can use RGB colors for distinguishing them. These in turn will increase KPIs.

The camera arrays are equipped with IR illuminators and exposure control to synchronize cameras and the IR-flash illuminators.

Figure 4: RGB-IR cameras

Processor chosen is the embedded system by NVIDIA: Jetson TX2 is the fastest, most power-efficient embedded AI computing device. This 7.5-watt supercomputer on a module brings AI computing with CUDA support (the basic AI languages that all DRAMA's modules are using). It's built around an NVIDIA Pascal™-family GPU and loaded with 8GB of memory and 59.7GB/s of memory bandwidth. The selection of NVIDIA embedded board allows the algorithms developed with Python/Tensorflow/CUDA to be easily portable between platforms (development and test environments).

Data capture and pre-annotation tool reuses the existing software from Smart Eye with modifications for this project purpose. The tools automate translation process from scenario scripts into timely instructions to test persons during the data capture sessions, categorizes the captured data into pre-defined classes of activities/behaviours and stores them in separate folders with labels.

Challenges: Lighting condition, mode switching between day and night-time, camera coverage, limited GPU capacity.

6.3 WP3 - Cabin monitoring and integration to driver monitoring

In this WP, we design machine learning algorithms to perform the following tasks in accordance with the selected scenarios in WP1 and the input data from the system developed in WP2:

- T3.1 - Driver pose/sitting position classification
- T3.2 - Driver/passenger activity classification
- T3.3 - Face expression classification
- T3.4 - Identification of person interaction within the cabin
- T3.5 - Supporting activity: Collecting and annotating data to train models

6.3.1 T3.1 - Driver pose/sitting position

Pose estimation has over the past years been greatly reshaped by CNN based approaches. DeepPose[17] was the first major paper that applied Deep Learning to Human pose estimation, outperformed traditional methods based on pictorial structures or deformable part models. The approach employs holistic reasoning to address the main challenges that besides extreme variability in articulations, many of the body joins are barely visible in the images.

Pose estimation algorithms for multi-person poses are categorized into 2 main classes:

- Top-down: Where the image is first analyzed to detect humans and then recognizing human pose in the detected regions (single pose recognition). Examples: DeepPose, Coarse-to-fine[18], Posefix[19].
- Bottom-up: All key points of different types of a human pose will be detected first and then joining algorithms will be applied to provide individual pose estimations. Examples: Openpose[20], Deepcut[21], PersonLab[22], [23], Densepose[24].

Of these categories, the bottom-up algorithms are preferred for this project because of simplicity and higher performance-cost ratio. Amongst these algorithms, the most active research developments are CMU's Openpose, Google's Personlab, and Facebook's Densepose. Openpose and Personlab are using 17 keypoint skeleton model: Nose, L/R Eyes, L/R Ears, L/R Shoulders, L/R Elbows, L/R Wrists, L/R Hips, L/R Knees, L/R Ankles. Densepose uses Skinned Multi-Person Linear model (SMPL) [25], a skinned vertex-based model that accurately represents a wide variety of body shapes in natural human poses.

We observe that the ankles are usually occluded from the camera images with the current camera settings, and the movement of lower part of human body does not create differences on the list of activities in focus. Therefore, in DRAMA prototypes, these key joints are not included in the skeleton model. These key joints and other key joints (such as hand skeleton) will be considered if different camera settings and activity list may require in future. Examples include in autonomous shuttle/bus interior designs where standing option in the vehicle is available.

The selected pose estimation algorithm for DRAMA project is Personlab with Mobilenet pretrained weights[26]. The motivation is that it provides modest accuracy[27], open license (Apache license), and the used joint skeleton model provides potentially more relevant information for activity recognition component than skin based SMPL model. Openpose can also be a good candidate in technical point of view, provided some restrictions by its current license.

A simple example usage of the Pose estimation module is the classification of whether a person is on or off recommended positions. With the assumption that the lower part of one's body is tighten to his/her seat by seatbelt, we propose a classification algorithm as follows:

- Collect the referenced key joint positions of the person at the recommended position (e.g. upright position with seatbelt on). The following key joints are selected: L/R eyes, nose, L/R ears, L/R shoulders.
- A body upper part rectangle (we name it PoseRect), representing on/off position of a person, is defined as the bounding box covering all above selected pose key joints.
- We compute PoseRects for the reference position ("On" position) and current position.
- Distance metrics between these two PoseRects is defined as 1 – "ratio between the intersection and the union area of the two rectangles". The distance is 1 when these two rectangles are complete matched and 0 if there is no overlapping. The distance captures 3D relative distance between the two positions.
- This distance is used for the classification of On/Off position by providing a threshold (e.g. 0.3)

6.3.2 T3.3 - Face expression recognition

Face expression recognition (FER) system includes major stages: image pre-processing, feature extraction and classification [28]:

- Preprocessing: This process improves the image qualities for the next step of feature extractions. Techniques include: cropping, scaling, contrast adjustment, normalization (e.g. histogram equalization) and most importantly localization (using e.g. Viola-Jones algorithm[29], [30] to detect the bounding boxed facial images from the input images).

- Feature extraction: The feature extraction methods are categorized into five types such as texture feature-based method, edge-based method, global and local feature-based method, geometric feature-based method and patch-based method.
- Classification: It is the final stage of FER system where computed feature vector of a face image is used to recognize the facial expressions. Paul Ekman and his colleagues investigated the production of emotional facial expressions and proposed that people display universal prototypical facial expressions that are specific to basic emotions[31], [32]: Happiness, Sadness, Fear, Disgust, Anger, Surprise and Neutral. This classification of 7 basic facial emotions has widely been accepted in this research area. Classification algorithms are available in wide variations: distance metric based [33], [34], KNN[35], SVM[34], [36]–[38], HMM[39], Decision tree, etc.

Convolution Neural Network (CNN), with its recent most success marked by Alexnet[40] in 2012 when it outperformed traditional image processing approaches the first time, provides the powerful tools for face expression recognition. However, the main issues of face expression recognition are on the subjectivity nature of the problem leading to unstable ground truth annotations, and "acting" emotions in the input datasets. Human's understanding of facial expressions varies with different cultures, living environments, and other experiences resulting in errors and bias of human annotations among different training datasets.

For the facial expression component within DRAMA system architecture, we have decided to use CNN approach with existing algorithms, since it is the latest approach with ongoing development in the field. This will help DRAMA system to benefit from latest developments, while not preventing it to use other methods as alternatives.

The training dataset used is Kaggle FER-2013 of 35000+ face expression images. The tested architectures include: Alexnet, ResNet and Xception. The selected architecture is Xception based on the small number of parameters and the recognition accuracy level (Accuracy obtained: 72.3%, # parameters 58,423). Figure 5 illustrates the network architecture of Xception and Figure 6 shows the preliminary results of the selected algorithms on Kaggle datasets together with our test person's images.

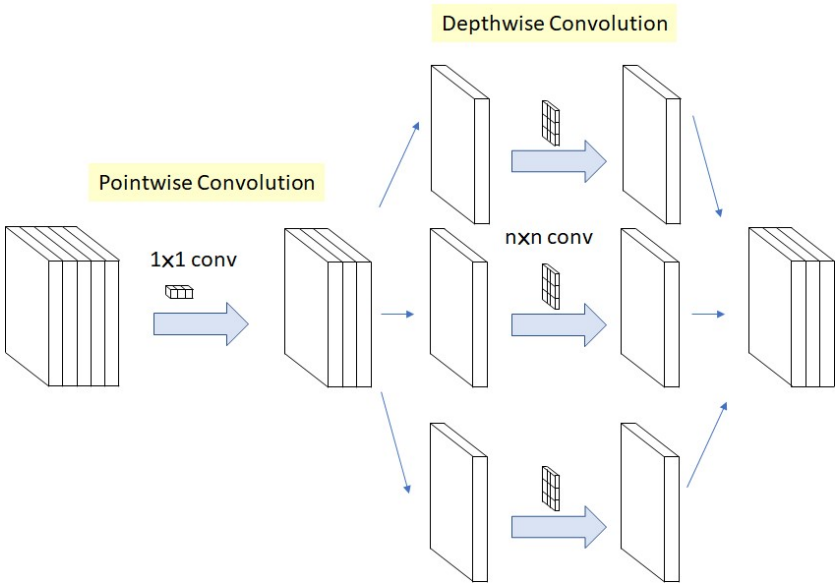


Figure 5 Xception network

Discussions:

The Ekman's suggestion that emotional expressions are universal stood largely unchallenged for a generation. But a new cohort of psychologists and cognitive scientists has been revisiting those data and questioning the conclusions. Researchers are increasingly split over the validity of Ekman's conclusions[41]. Within this task, we observe that the datasets are influenced by "acting" expressions, and the low accuracy on the emotion recognition. Since the main objective of DRAMA system is not exactly the labelled face expressions, we believe that providing more related information about face expressions to the next modules in pipeline will still help to provide good results for safety/conform despite the fact that the classification of expressions is not at high accuracy. Therefore, we also provide face landmark as part of the feature vector.

Figure 6: Face expression recognition using retrained Xception CNN network

6.3.3 T3.2/3.4 - Driver/passenger activity and interaction recognition

6.3.3.1 Algorithm highlevel design

We first conduct a literature review on the state of the art algorithms for human activity recognition and group activity recognition [42]–[65].

Depending on complexity levels, human activities are categorized as proposed by [45] into: (i) gestures[66]; (ii) atomic actions [67]; (iii) interactions[68]; (iv) group actions [69]; (v) behaviours [70]; and (vi) events[71].

Within the scope or DRAMA project, we set our focus of task T3.2 on atomic actions/activities and T3.4 on interactions. The extracted activity feature vectors (by design) contain also crucial information for in cabin recognition of gestures, behaviours, and events.

Objects play a valuable role in human activity recognition since they can provide strong cues about the involved actions. Recent researches have shown that combined activity recognition and object detection are mutually beneficial comparing to addressing these problems separately[72], [73]. Cognitive psychological studies (e.g.[74]) also highlighted that intentional actions always connect to desired goals, resulting in object type and movements information is crucial for interaction recognition as they are tightly connected to the goals of the interactions involving objects.

Challenges:

- Inadequacy of available training/validation datasets: The datasets were created for specific activities, mainly outdoor activities.
- Publicly available algorithms are specific to datasets and problems to be solved.
- Algorithms are not always designed for real-time requirement.
- A family of algorithms relies on depth information of visual data, which adds higher technical requirements for the data acquisition systems (3D depth cameras and more computing resources).
- The recognition of interactions between humans and objects has a high level of semantic information. More semantic focused approaches should be provided
- Localization (temporal, spatial) of activity/interaction is still open research question.

The target environment of DRAMA, i.e. car interior, has the following properties:

- Known background.
- Unchanged (or minimal changed) of human/object existence and quantity found inside the cabin along a single journey.
- Input data are provided in form of video streams.

Therefore, the below assumptions are hold and used for the design of project's algorithms:

- Movements detected by fix-mounted cameras are highly connected to activities, not to background changes in image sequences
- Body pose can be located to seat position
- Size, distance and location constancy is simpler.

Figure 7 illustrates the logical architecture of the DRAMA framework. At the top level, 5 recognition modules are selected based on the following criteria:

- Publicly availability or existing modules/products.
- Provide independent perspectives that may potentially provide cue information for the recognition of activities/interactions. Spatiotemporal relations between the selected sub-features will compose the "activity feature vector", a fixed length vector containing information about activities/interactions.

Based on this observation, we select (and customize/retrain) the state-of-the-art modules to capture the followings from the video sequences:

- Position/Pose of person involved in an activity.
- Object detection and recognition (object type, size and relative location).
- Movement of people and objects (optical flow).

Finally, recurrent neural network component is added to capture and collect detected sequential action cues from the timeseries of activity feature vectors. In combination with other components (e.g. face expressions), we can also compose a feature vector representing group activities (e.g. conversation, arguments).

The logical architecture of the algorithm is provided in Figure 7. The top layers are elementary (mapping) modules, the middle layer contains activities and interactions and the bottom layer contains group activities/behavior.

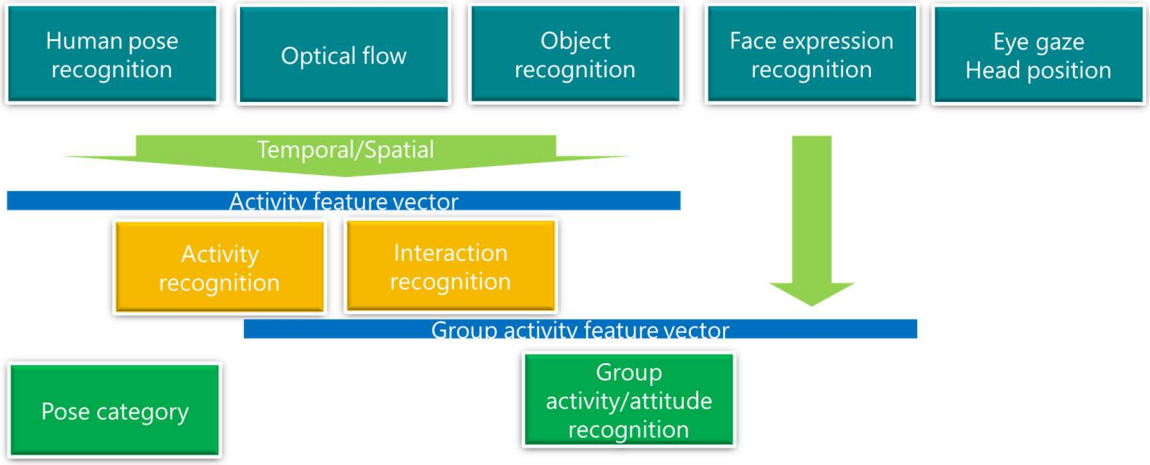


Figure 7: Logical architecture of DRAMA

6.3.3.2 Algorithm detailed design

The proposed scheme of the activity/interaction recognition part of DRAMA system is illustrated in Figure 8.

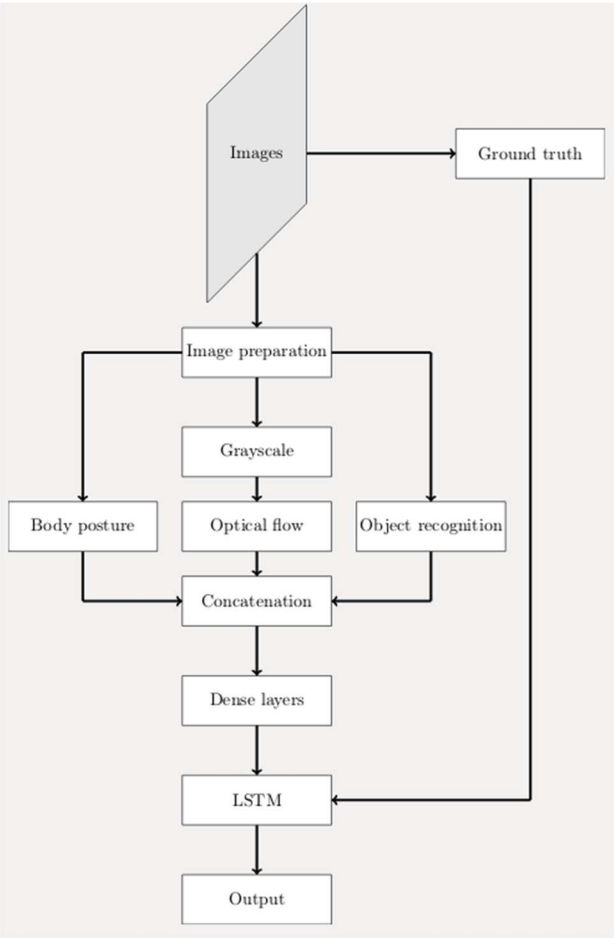


Figure 8; Schematic overview of action classification algorithm

6.3.3.2.1 Pre-processing

The captured sequences of RGB/-IR images are converted into single-channel (grayscale) images if needed. In RGB-IR cameras, this pre-processing step is equivalent to extraction of IR images during night-time and RGB2Gray during daytime. The idea of using single-channel images will help the architecture being less sensitive to illumination changes. The use of IR illuminator as controlled light source can also reduce the external illumination source impacts.

The input images captured from different cameras at the same timestamp are also stitched to create a wider view single image. Fisheye distortion correction is also applied to remove the fisheye distortion in the input images.

A simple camera calibration using aruco markers [75] is also used in our experiments, see Figure 9.

Figure 9: Aruco markers used to align input images

Human detection is applied to facilitate single pose recognition module (if applicable) and to provide seat occupancy information. This detection can accommodate activity scenario of "forgotten child in car", a special activity where human appears not as actor but subject.

6.3.3.2.2 Activity sub-feature extraction

These prepared images are then fed to the three parallel processing streams: (i) body posture recognition, (ii) object recognition and (iii) optical flow fields. The body posture recognition provides skeleton model estimations of all found persons in the cabin (from multiple input cameras). The object recognition module provides object classification and tracked relative location for objects falling into the selected classes of interest. Optical flow field captures the local movements of image pixels. Since the cameras are stationary relatively to the car cabin, these local movements are assumed to be highly related to the movements resulted by in-cabin activities.

We selected PoseNet [23] for body posture recognition, which creates multi-person skeletal models based on 2D coordinates of body joints. Another possible alternative is OpenPose [20], which may also include hand gesture recognition. The normalized vector forming from the 2D coordinates of 13 detected joints (excluding knees and ankles as they are usually occluded from the images) are then used as pose sub-feature vector. The normalization is performed to remove the dependency on absolute position of the skeleton in the image and provides the scale constancy.

The features extracted by optical flow capture the short-term temporal context in the sequence data and is calculated with the Farneback algorithm [76] for consecutive image pairs. Trade-off between time and accuracy can also be achieved by selected image pairs of different time-gap scales. The Farneback algorithm is used to create dense optical flow fields. The flow fields are then segmented into 10 partly overlapping rectangle shaped

areas, where each area will be represented by the histogram of the flow vector angles created inside. Combination of these area representation creates optical flow sub-feature vector. This optical flow sub-feature vector as such captures short-term temporal movements. The dense optical flow results in sub-feature vectors of constant length.

For the object recognition module, the state of the art algorithm is YOLOv3 [77]. Since the in-cabin environment is predictable, we also consider another alternative using regions of interest (RoI) based implementation of mobilenetv2 [26]. The input images are segmented into several RoI's and the module predicts the existence of objects inside each region. Recognized object classes and positions can provide valuable input for action recognition that involve objects. This could also provide safety-related information e.g. objects laying around that can cause damage or injury in some driving situations. Both alternatives are lightweight and can provide real-time performance. We also apply tracker function using Kalman filter[78] to increase the visibility of object traces. An object sub-feature vector is then created as an array of objects ordered by class ID, together with position and detection accuracy.

The normalized sub-feature vectors extracted from these three modules at each time stamp are then concatenated to create an "activity" feature vector. This "activity" feature vector is considered as a snapshot of all activity-related information at a specific timestamp. The concatenation operator enables flexibility in adopting new or modified sub-feature extraction modules. Timeseries of the feature vectors will include temporal relationship information to define more complex actions/behaviours. The sequences of activity feature vectors are then used as input to a set of dense neural network layers, further reducing the dimension.

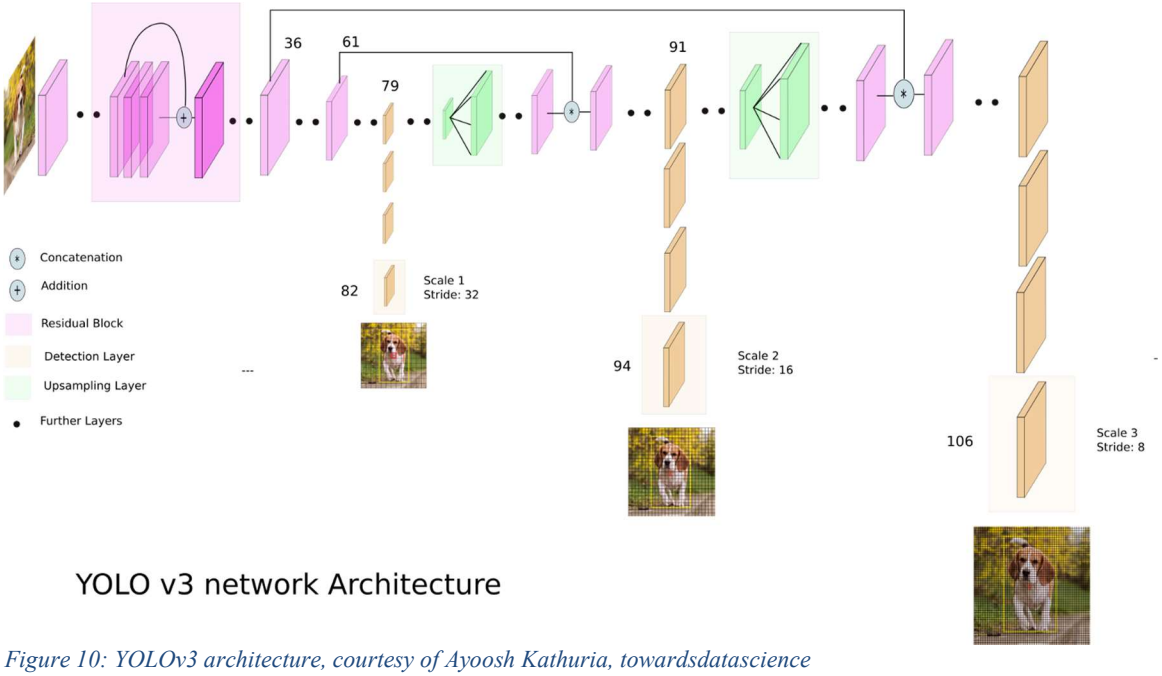


Figure 10: YOLOv3 architecture, courtesy of Ayoosh Kathuria, towardsdatascience

6.3.3.2.3 Activity classification

The activity classification module consists of 2 main components:

- Stacked fully connected blocks, where each block consists of a dense layer following by Batchnorm [79], leaky ReLU [80] and dropout[81] layers respectively.
- LSTM[82] layers, acting as longer terms temporal relation information extractor. Future research ideas may also include a more complex structure of LSTM layers to capture also non-sequential relationships.

Once the individual actions have been classified for each of the cameras at a specific time point the interactions can be determined. The LSTM gives a classification separately for each of the cameras where the interactions are one of the possible classes. The predictions from different cameras are merged and compared to find matching interactions. With this structure, we can use the same network for different cameras.

Following steps are taken as normalization methods:

- Use relative positions rather than absolute. Reference points include the detected aruco markers and body pose center point.
- Normalize image based distances between points to the scale of (0..1).
- Normalize of sub-feature vectors by L2 norm.

6.3.4 T3.5 - Data collection

There is unfortunately no publicly available dataset for training of driver and passenger in-vehicle activities. We conducted six iterations of data capture and annotation for the designed system. Four iterations for calibrating different feature extraction methods and two for activity recognition training/validation.

Data capture iterations and objectives are listed in below table:

No	Setup	Objectives
1	Small amount of data, office environment, one participant.	Calibrate the cameras and data collection tools
2	Office environment, two test persons interacting with objects.	proof test and calibrate object and pose recognition sub-modules.
3	Selected objects, office/in-cabin simulation environment.	Object data capture to retrain object recognition module.
4	Car environment, two test participants (driver and front passenger seats).	Calibrate the camera mounting configurations in the car and preliminarily validate object detection and body posture recognition module in car cabin environment.

5	Car environment, two test participants (driver and front passenger seats).	Preliminary dataset for actions to proof test the activity recognition architecture
6	Car environment, 2 test participants (driver, front and back passenger seats).	A larger dataset was needed with more data and more scenarios in order to better validate the results, while reducing unwanted effects that could otherwise increase the accuracy at the cost of generality such as overfitting

Captured videos were classified and annotated by camera, action and position.

Figure 11: Data capture, in-cabin environment

Figure 12: Data capture, in cabin environment

6.4 WP4 - Scenario evaluation

6.4.1 Algorithm performance evaluation

The performance evaluation of the system consists of evaluations of all building blocks and the entire system. The main metrices used for evaluation is accuracy, recall and precision. Prototypes of the complete model were created to run in real time.

Evaluations on the contribution of each sub-feature modules into the final action recognition results have been performed by comparing the prototypes using only any 2 out of 3 sub-features with the one that uses all 3 (Figure 13 shows confusion matrices of these experiments where body pose sub-feature is taken into account). Performance metrics of sub-modules are taken from the reported evaluations.

A post processing step was implemented where rule-based corrections were made to recognized actions based on what objects were detected and if the other cameras detected interactions. The system was also tested with and without this post processing step to get further insights into how the object detection influence the classifications and the ability of using multiple cameras to detect interactions.

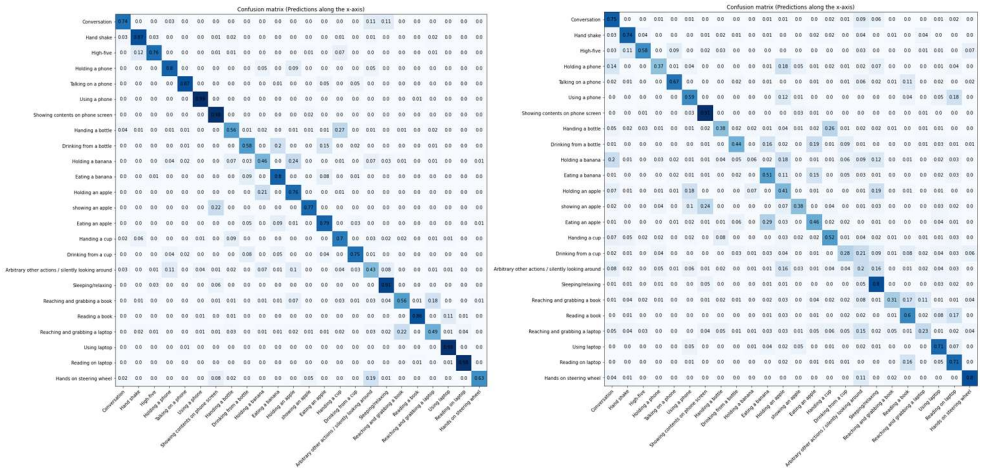


Figure 13: Confusion matrix of the activity recognition system with (left) and without (right) body pose sub-feature (plotted with matplotlib[83])

Evaluations also performed to compare between using absolute and relative coordinates in different sub-feature modules.

6.4.2 Scenario evaluation

We apply Software Architecture Analysis Method (SAAM)[84] for scenario evaluation. SAAM proposes five step approach to analyse how well a software architecture can adapt to new requirements.

6.4.2.1 Develop scenarios

Scope	Description
Users	- OEM's and/or car manufacturers that can use the system predicted outputs to provide safety/comfort improvement system. - Researchers/developers want to contribute to the algorithms.

Usages	- Data capture and system re-/training
	- Incorporation of newly arrived requirements:
	  o New object type
	  o New passenger status

Scope	Description
	○ New type of passenger activity/interaction ○ Integration of new sensors (more cameras, changing of mounting positions) - Real-time recognitions of: Pose, Seat occupancy, driver/passenger activities, face expression

6.4.2.2 Describe architecture

The algorithm architecture is design with modular and hierarchy. This structure is used to accommodate recognition of foreseen safety/comfort related activities. Reader consults Section 6.3 for more details.

6.4.2.3 Classify and prioritize scenarios

Selected scenarios and priority are provided in Annex 10.1.

The indirect scenarios are derived from foreseen possible future requirements of the system:

- Adding/removing sub-feature modules.
- Adding/removing classes of object, activity, expression, pose
- Adding/removing training samples of specific objects, poses, activities
- Changing the tradeoff balance between accuracy and system realtime performance
- Adding/removing sensor types (motion detection, proactive lighting, control signals from cars)
- Changing environment

6.4.2.4 Individually evaluate indirect scenarios

Evaluations were performed against different setup as listed in Section 6.4.2.3. The modular design of the architecture helps to maintain linearity of the algorithm's complexity when adding indirect scenarios. Evaluations of accuracy gains were performed by comparing confusion matrices with different indirect scenarios added (same convergent criteria).

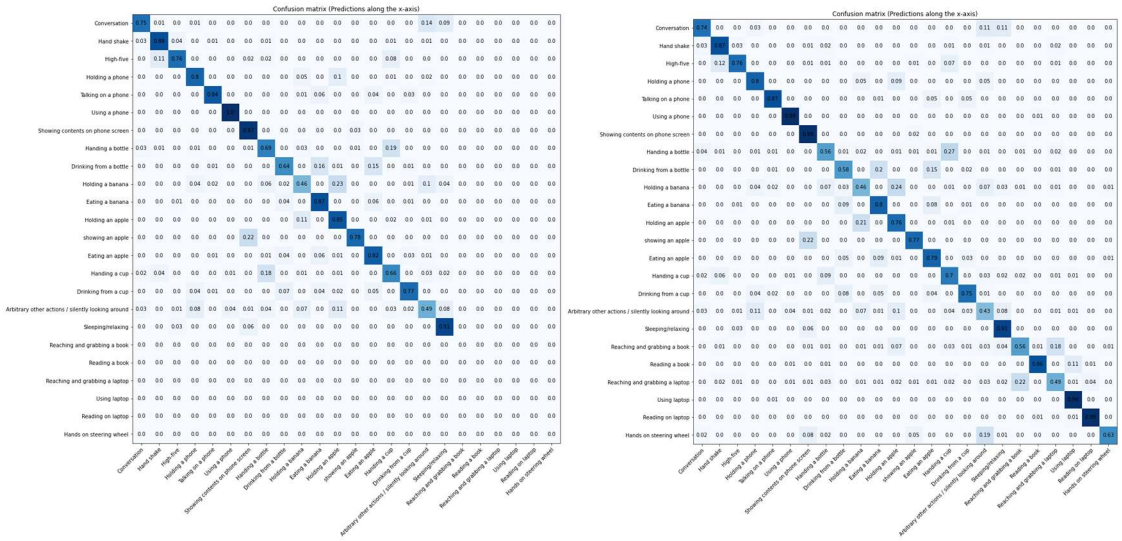


Figure 14: Confusion matrices with adding/removing activity classes (plotted with matplotlib[83])

6.4.2.5 Assess scenario interaction

The algorithm architecture does not allow two or more scenarios are requesting changes over the same component(s). The accuracies of different system's outputs can thus be improved by improving different modules (new module algorithm or newly trained weights) without risk of degradation of existing scenario's accuracies.

6.4.2.6 Create an overall evaluation

The algorithm design does not enforce any priority rule over different information dimension of its outputs. In fact, the priority order should be optimized with regards to the specific requirements of the system's usages, i.e. safety/comfort applications. By proposing this architecture, our argument is that it provides conceptually all potential important information extracted from input sensors that can be used for human activity/behaviour recognition. The selected sub-feature vectors are uncorrelated and represent different dimensions of in-cabin activities. The same design principles can be applied when more input sensor types or more outputs may be required in future (e.g. blood pressure, temperature sensors, recognition of in-cabin context…).

The evaluations within this project are performed based on the selected activity scenarios as discussed in Section 6.1 where all activities are considered equally important. Figure 15 displays confusion matrix of the recognition results classifying 24 activities of interest.

More evaluations can be performed in accordance with specific use cases and requirements as provided by industrial applications regarding safety/comfort. The evaluation metrics will be then tailored to the specific requirements to optimize the solution accordingly. We will therefore propose new research ideas to apply the algorithm developed in DRAMA for estimation of different parameters (e.g. driver re-engagement time) based on the recognized on-going in-cabin activity context.

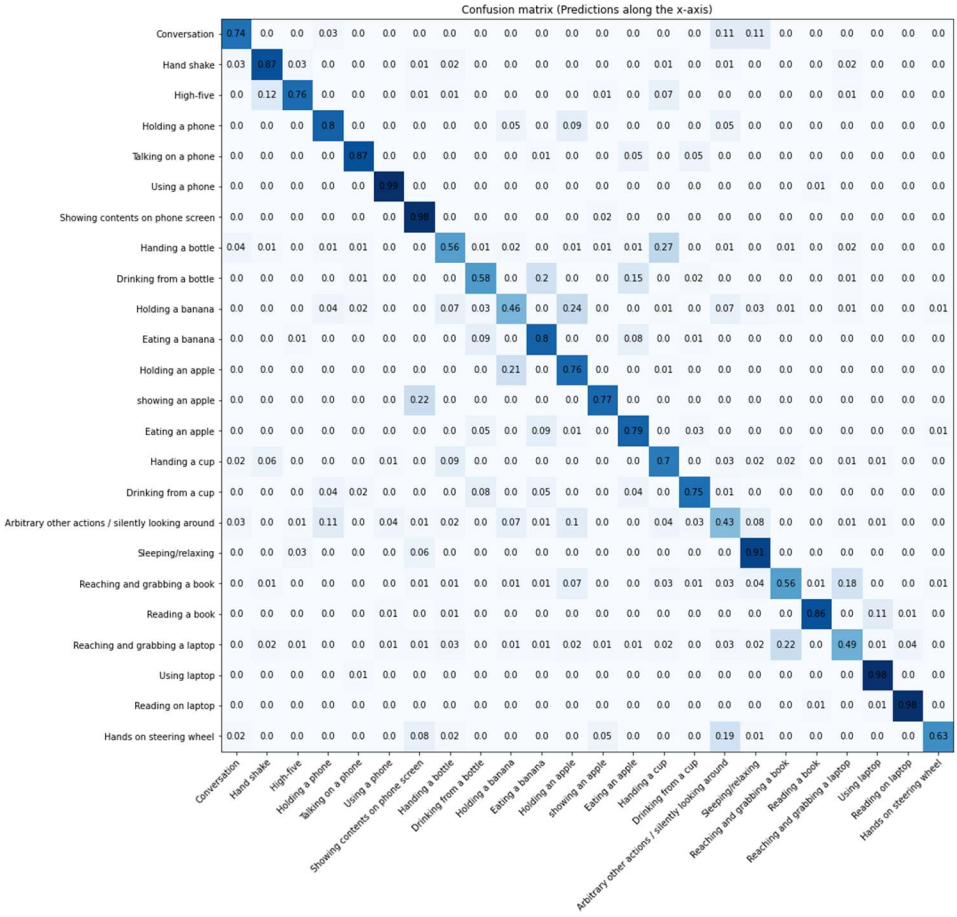


Figure 15: Activity recognition performance (plotted with matplotlib[83])

6.5 Prototypes

Three prototypes have been developed to proof the proposed concepts. The information flow of the prototypes is provided in Figure 16. Recognition results of each sub-feature are classified by seat position. Information from all streams will provide inputs for activity recognition layers. The fusion of results from all modules are used for the presentation layer. For the demonstration purpose, representation layer is prototyped with third-person view (prototypes 1,3) and first-person view (prototype 2). In the actual application, this presentation layer will be part of the applications that use DRAMA system as input.

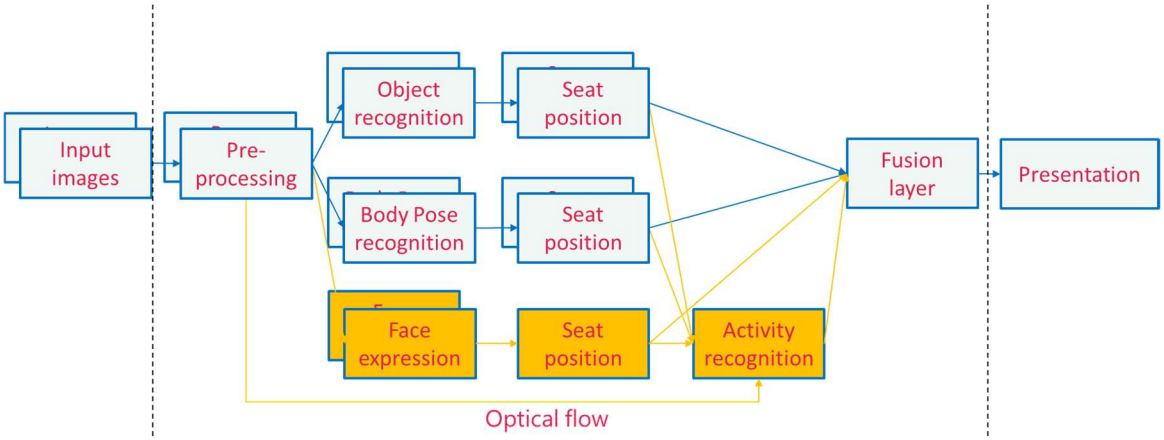


Figure 16: DRAMA information flow

6.5.1 Prototype 1

The system in prototype 1 uses car cabin setup (Figure 12) with 3 RGB webcams capturing driver, front-seat passenger and back-seat passengers.

The prototype extracts body pose (17 key-joints skeleton model), face landmarks (68 feature points), facial expressions, dense optical flow and object detection + tracking, and recognize the activity. All recognized information is overlaid in the representation screen as shown in Figure 17.

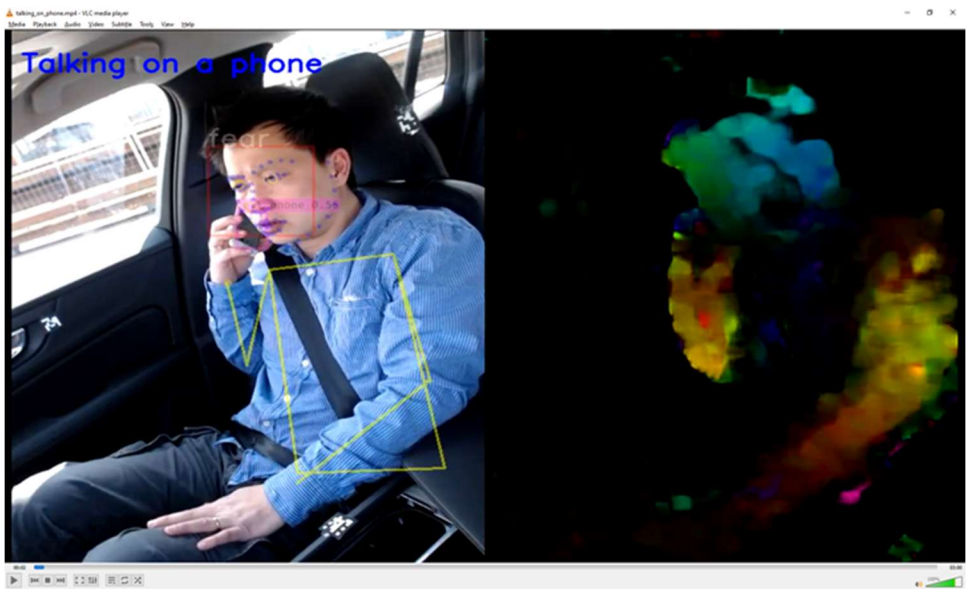


Figure 17: Prototype 1 - HAR in cabin

6.5.2 Prototype 2

Prototype 2 is a car simulator, setup in the office space (Figure 18). The prototype applies all ideas with retrained components to demonstrates some application scenarios including:

- Object detection + Placement
  - Person (+ Seat occupancy)
  - Cup
  - Apple
  - Phone
- Body pose recognition
- Action recognition
  - Holding phone
  - Talking on the phone
  - Holding cup
  - Drinking
  - Holding apple
  - Eating
- Facial expressions

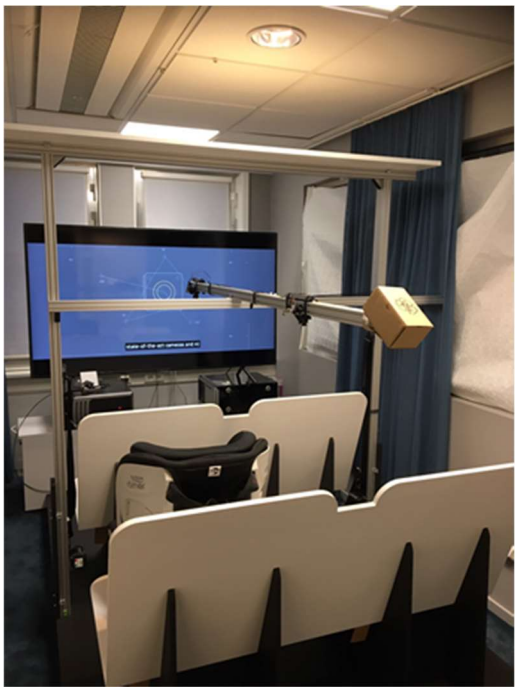


Figure 18: DRAMA prototype 2, simulation room

The prototype 2's screen (Figure 19) shows symbols representing different recognition results in real-time.

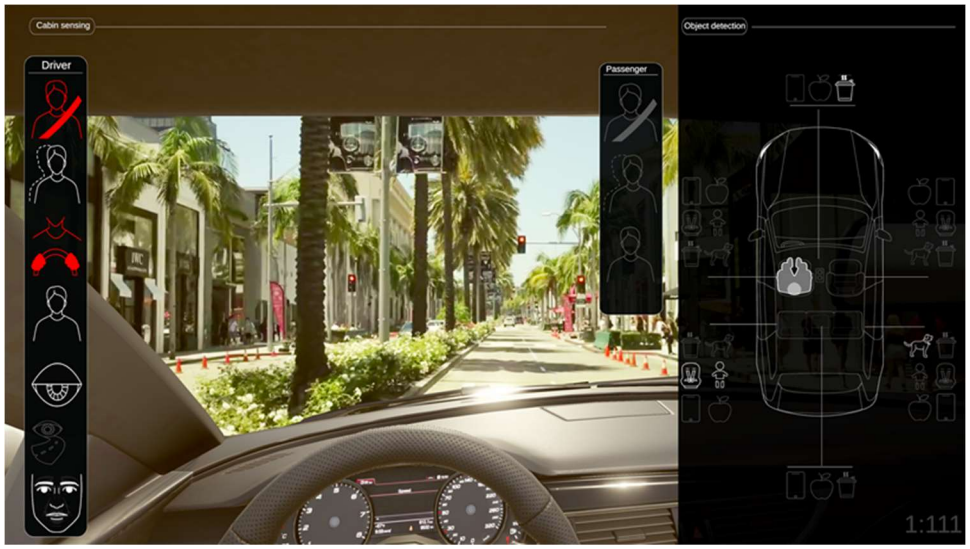


Figure 19: Screenshot of simulation room windshield

6.5.3 Prototype 3

Prototype 3 was described earlier in chapter 6.2. Different HW and SW settings were tested and used for data collection.

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	The project helps the consortium in gaining broad and deep knowledge on interior sensing. The introduction of new product line and project dissemination will also increase knowledge and awareness of other actors in the ecosystem. This will pave the way to the future system where better understandings between both insides and outsides of the vehicles can be combined to create a safer and more pleasure journeys in highly automated vehicles.

How are the project results planned to be used and disseminated?	Mark with X	Comment
Be passed on to other advanced technological development projects	X	Project results in several research ideas to be further exploited: - Data capture and annotation automation - Application to estimate distraction and re-engagement of drivers in SAE3 vehicles - Inclusion of wearable and health related devices for applications e.g. mobility for all.
Be passed on to product development projects	X	Project results are used as part of new Smart Eye interior sensing product line [85]
Introduced on the market	X	Prototype and the product line were introduced in CES 2020
Used in investigations / regulatory / licensing / political decisions		Not yet, but could be in the future.

7.2 Publications

[1] M. Torstensson, B. Duran, and C. Englund, "Using Recurrent Neural Networks for Action and Intention Recognition of Car Drivers," presented at the 8th International Conference on Pattern Recognition Applications and Methods, 2020, pp. 232–242.
[2] M. Torstensson, T. H. Bui, D. Lindström, C. Englund, and B. Duran, "In-vehicle Driver and Passenger Activity Recognition," presented at the Third Swedish Symposium on Deep Learning, 2019, p. 4.
[3] T. Wilhelm, "Towards Facial Expression Analysis in a Driver Assistance System," in 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), 2019, pp. 1–4, doi: 10.1109/FG.2019.8756565.

8 Conclusions and future research

The aim of the project was to take the first step to describe how computer vision and machine learning techniques can be used for automatically recognize in-cabin

activities/situations. The process gave lots of insights about what are most common activities inside SAE3-5 car cabin, how to recognize them in real-time and how to implement such system in car.
Based on the literature and SoA of available algorithms, the project proposes a general framework and structural approach to decompose in-cabin human activities into conceptual atomic components that can then be accommodated by existing and evolving algorithms. This structure also creates the first step towards explainable design of the machine learning algorithms for automotive industry. Several prototypes have also been implemented to prove the proposed concept, that attract attentions from industry.
The key question is obviously the robustness and accuracy of the system in compliance to safety standards in vehicle industry. This opens the opportunities for future researches to investigate practical safety/comfort applications as required by the industry, and how one can implement such system in compliance with functional safety standards.

9 Participating parties and contact persons

RISE Research Institutes of Sweden AB
Lindholmspiren 3A
417 56 Göteborg
Contact: Thanh Hai Bui
thanh.bui@ri.se

Smart Eye Aktiebolag
Första Långatan 28 vån 7,
413 27 Göteborg
Contact: Henrik Lind
Henrik.lind@smarteye.se

References

[1] J. Stewart, “Self-Driving Cars Won’t Just Watch the Road. They’ll Watch You, Too,” *Wired*, Feb. 13, 2017.
[2] T. E. Dokor, “Autonomous Vehicles Need In-Cabin Cameras to Monitor Drivers,” *IEEE Spectrum: Technology, Engineering, and Science News*, Apr. 10, 2016. https://spectrum.ieee.org/cars-that-think/transportation/self-driving/autonomous-vehicles-need-incabin-cameras-to-monitor-drivers (accessed May 10, 2019).

[3] D. J. McDuff, E. B. Blackford, and J. R. Estepp, “Fusing Partial Camera Signals for Noncontact Pulse Rate Variability Measurement,” *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 8, pp. 1725–1739, Aug. 2018, doi: 10.1109/TBME.2017.2771518.
[4] N. Arbabzadeh and M. Jafari, “A Data-Driven Approach for Driving Safety Risk Prediction Using Driver Behavior and Roadway Information Data,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 2, pp. 446–460, Feb. 2018, doi: 10.1109/TITS.2017.2700869.
[5] J. A. Michon, “A Critical View of Driver Behavior Models: What Do We Know, What Should We Do?,” in *Human Behavior and Traffic Safety*, L. Evans and R. C. Schwing, Eds. Boston, MA: Springer US, 1985, pp. 485–524.
[6] G. Castignani, T. Derrmann, R. Frank, and T. Engel, “Driver Behavior Profiling Using Smartphones: A Low-Cost Platform for Driver Monitoring,” *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 1, pp. 91–102, Spring 2015, doi: 10.1109/MITS.2014.2328673.
[7] “Ambarella and Smart Eye partner to deliver next generation AI-based Driver Monitoring,” *News Powered by Cision*. https://news.cision.com/smart-eye/r/ambarella-and-smart-eye-partner-to-deliver-next-generation-ai-based-driver-monitoring,c2711299 (accessed Feb. 10, 2020).
[8] T. W. Victor, M. Dozza, J. Bärgman, C.-N. \AAkerberg Boda, J. Engström, and G. Markkula, “Analysis of Naturalistic Driving Study Data: Safer Glances, Driver Inattention, and Crash Risk,” 2014, doi: 10.17226/22297.
[9] T. R. Board, National Academies of Sciences Engineering, and Medicine, *Design of the In-Vehicle Driving Behavior and Crash Risk Study*. Washington, DC: The National Academies Press, 2011.
[10] J. S. Hickman, R. J. Hanowski, and J. Bocanegra, *Distraction in Commercial Trucks and Buses: Assessing Prevalence and Risk in Conjunction with Crashes and Near-Crashes*. 2010.
[11] T. A. Dingus *et al.*, “Driver crash risk factors and prevalence evaluation using naturalistic driving data,” *PNAS*, vol. 113, no. 10, pp. 2636–2641, Mar. 2016, doi: 10.1073/pnas.1513271113.
[12] Ipsos/GenPop, “What is the future,” *Spring 2018*. https://future.ipsos.com/category/wtf/spring-2018 (accessed Feb. 10, 2020).
[13] S. Jorlöv, K. Bohman, and A. Larsson, “Seating Positions and Activities in Highly Automated Cars – A Qualitative Study of Future Automated Driving Scenarios,” in *IRCOBI Conference Proceedings*, 2017, Accessed: May 10, 2019. [Online]. Available: https://trid.trb.org/view/1485922.
[14] “Will Self-Driving Cars Eliminate Distracted Driving? Most People Think So, Says New National Survey,” *erieinsurance.com*. http://www.erieinsurance.com/blog/distracted-driving-survey (accessed Feb. 10, 2020).
[15] S. Writer, “What would you do in a self-driving car?” https://engineering.cmu.edu/news-events/news/2017/01/25-av-car-survey.html (accessed Feb. 10, 2020).

[16] S. Das, A. Sekar, R. Chen, H. C. Kim, T. J. Wallington, and E. Williams, "Impacts of Autonomous Vehicles on Consumers Time-Use Patterns," *Challenges*, vol. 8, no. 2, p. 32, Dec. 2017, doi: 10.3390/challe8020032.
[17] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, Jun. 2014, doi: 10.1109/CVPR.2014.214.
[18] U. Iqbal and J. Gall, "Multi-person Pose Estimation with Local Joint-to-Person Associations," in *Computer Vision – ECCV 2016 Workshops*, Cham, 2016, pp. 627–642, doi: 10.1007/978-3-319-48881-3_44.
[19] G. Moon, J. Y. Chang, and K. M. Lee, "PoseFix: Model-agnostic General Human Pose Refinement Network," *arXiv:1812.03595 [cs]*, Mar. 2019, Accessed: Feb. 14, 2020. [Online]. Available: http://arxiv.org/abs/1812.03595.
[20] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *arXiv:1812.08008 [cs]*, Dec. 2018, Accessed: May 10, 2019. [Online]. Available: http://arxiv.org/abs/1812.08008.
[21] L. Pishchulin *et al.*, "DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation," *arXiv:1511.06645 [cs]*, Apr. 2016, Accessed: Feb. 14, 2020. [Online]. Available: http://arxiv.org/abs/1511.06645.
[22] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model," in *Computer Vision – ECCV 2018*, Cham, 2018, pp. 282–299, doi: 10.1007/978-3-030-01264-9_17.
[23] R. Wightman, *A Python port of Google TensorFlow.js PoseNet (Real-time Human Pose Estimation): rwightman/posenet-python*. 2019.
[24] Y. Zhou, J. Deng, and S. Zafeiriou, "Improve Accurate Pose Alignment and Action Localization by Dense Pose Estimation," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 480–484, doi: 10.1109/FG.2018.00077.
[25] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A Skinned Multi-Person Linear Model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.
[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *arXiv:1801.04381 [cs]*, Jan. 2018, Accessed: May 10, 2019. [Online]. Available: http://arxiv.org/abs/1801.04381.
[27] "COCO - Common Objects in Context." http://cocodataset.org/index.htm#keypoints-leaderboard (accessed Aug. 08, 2019).
[28] "A Survey on Human Face Expression Recognition Techniques | Elsevier Enhanced Reader." https://reader.elsevier.com/reader/sd/pii/S1319157818303379?token=4EFFDA4A321CC2C1B1A8CC3714598EDCB3803673BC0CF92DF5CDCB65F1AEA0059E5B22605AD852310AC1C4E0A993AA38 (accessed Feb. 14, 2020).
[29] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 2001, vol. 1, pp. I-511-I–518, doi: 10.1109/CVPR.2001.990517.
[30] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004, doi: 10.1023/B:VISI.0000013087.49260.fb.
[31] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *Nebraska Symposium on Motivation*, vol. 19, pp. 207–283, 1971.
[32] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J Pers Soc Psychol*, vol. 17, no. 2, pp. 124–129, Feb. 1971, doi: 10.1037/h0030377.
[33] Y. Gao, M. K. H. Leung, S. C. Hui, and M. W. Tananda, "Facial Expression Recognition from Line-Based Caricatures," *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, vol. 33, no. 3, pp. 407–412, 2003, doi: 10.1109/TSMCA.2003.817057.
[34] G. P. Hegde, M. Seetha, and N. Hegde, "Kernel Locality Preserving Symmetrical Weighted Fisher Discriminant Analysis based subspace approach for expression recognition," *Engineering Science and Technology, an International Journal*, vol. 19, no. 3, pp. 1321–1333, Sep. 2016, doi: 10.1016/j.jestch.2016.03.005.
[35] A. Poursaberi, H. A. Noubari, M. Gavrilova, and S. Y. Yanushkevich, "Gauss-Laguerre wavelet textural feature fusion with geometrical information for facial expression identification," *EURASIP J. Image Video Process.*, vol. 17, no. 1, pp. 1–13, 2012.
[36] G. Zhao and M. Pietikäinen, "Boosted multi-resolution spatiotemporal descriptors for facial expression recognition," *Pattern Recognition Letters*, vol. 30, no. 12, pp. 1117–1127, Sep. 2009, doi: 10.1016/j.patrec.2009.03.018.
[37] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, Jan. 2015, doi: 10.1109/TAFFC.2014.2386334.
[38] S. Kumar, M. K. Bhuyan, and B. K. Chakraborty, "Extraction of informative regions of a face for facial expression recognition," *IET Computer Vision*, vol. 10, no. 6, pp. 567–576, 2016, doi: 10.1049/iet-cvi.2015.0273.
[39] M. H. Siddiqi, R. Ali, A. Sattar, A. M. Khan, and S. Lee, "Depth Camera-Based Facial Expression Recognition System Using Multilayer Scheme," *IETE Technical Review*, vol. 31, no. 4, pp. 277–286, Jul. 2014, doi: 10.1080/02564602.2014.944588.
[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[41] D. Heaven, "Why faces don't always tell the truth about feelings," *Nature*, vol. 578, no. 7796, pp. 502–504, Feb. 2020, doi: 10.1038/d41586-020-00507-5.
[42] J. K. Aggarwal and M. S. Ryoo, *Human Activity Analysis: A Review*. .
[43] X. Li and M. C. Chuah, "ReHAR: Robust and Efficient Human Activity Recognition," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, pp. 362–371, doi: 10.1109/WACV.2018.00046.
[44] K. Stephens, "Human and Group Activity Recognition from Video Sequences," phd, University of York, 2016.

[45] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, “A Review of Human Activity Recognition Methods,” *Front. Robot. AI*, vol. 2, 2015, doi: 10.3389/frobt.2015.00028.
[46] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, “A Review on Human Activity Recognition Using Vision-Based Method,” *J Healthc Eng*, vol. 2017, p. 3090343, 2017, doi: 10.1155/2017/3090343.
[47] 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)*, Mar. 2018, pp. 1625–1632, doi: 10.1109/WACV.2018.00180.
[48] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, “A Hierarchical Deep Temporal Model for Group Activity Recognition,” *CoRR*, vol. abs/1511.06040, 2015, Accessed: May 22, 2019. [Online]. Available: http://arxiv.org/abs/1511.06040.
[49] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, “Recurrent Neural Networks for driver activity anticipation via sensory-fusion architecture,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 3118–3125, doi: 10.1109/ICRA.2016.7487478.
[50] A. Manzi, L. Fiorini, R. Limosani, P. Dario, and F. Cavallo, “Two-person activity recognition using skeleton data,” *IET Computer Vision*, vol. 12, no. 1, pp. 27–35, Sep. 2017, doi: 10.1049/iet-cvi.2017.0118.
[51] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, “Deep Learning for Sensor-based Activity Recognition: A Survey,” *Pattern Recognition Letters*, vol. 119, pp. 3–11, Mar. 2019, doi: 10.1016/j.patrec.2018.02.010.
[52] J. A. Ward, P. Lukowicz, and H. W. Gellersen, “Performance Metrics for Activity Recognition,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 6:1–6:23, Jan. 2011, doi: 10.1145/1889681.1889687.
[53] M. Asadi-Aghbolaghi *et al.*, “A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences,” in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 476–483, doi: 10.1109/FG.2017.150.
[54] L. Wang, Y. Qiao, and X. Tang, “Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4305–4314, Jun. 2015, doi: 10.1109/CVPR.2015.7299059.
[55] H. S. Koppula, R. Gupta, and A. Saxena, “Learning human activities and object affordances from RGB-D videos,” *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, Jul. 2013, doi: 10.1177/0278364913478446.
[56] S. Lathuilière, G. Evangelidis, and R. Horaud, “Recognition of Group Activities in Videos Based on Single-and Two-Person Descriptors,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2017, pp. 217–225, doi: 10.1109/WACV.2017.31.
[57] M. S. Ryoo, C.-C. Chen, J. K. Aggarwal, and A. Roy-Chowdhury, “An Overview of Contest on Semantic Description of Human Activities (SDHA) 2010,” in *Recognizing Patterns in Signals, Speech, Images and Videos*, 2010, pp. 270–285.

[58] H. Kuehne, A. Arslan, and T. Serre, “The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, Jun. 2014, pp. 780–787, doi: 10.1109/CVPR.2014.105.

[59] J. Bütepage, M. Black, D. Kragic, and H. Kjellström, “Deep representation learning for human motion prediction and classification,” Feb. 2017, Accessed: May 10, 2019. [Online]. Available: https://arxiv.org/abs/1702.07486v2.

[60] S. Herath, M. Harandi, and F. Porikli, “Going Deeper into Action Recognition: A Survey,” *arXiv:1605.04988 [cs]*, May 2016, Accessed: May 10, 2019. [Online]. Available: http://arxiv.org/abs/1605.04988.

[61] Z. Gharaee, P. Gärdenfors, and M. Johnsson, “Online recognition of actions involving objects,” *Biologically Inspired Cognitive Architectures*, vol. 22, pp. 10–19, Oct. 2017, doi: 10.1016/j.bica.2017.09.007.

[62] T. Hao, D. Wu, Q. Wang, and J.-S. Sun, “Multi-view representation learning for multi-view action recognition,” *Journal of Visual Communication and Image Representation*, vol. 48, pp. 453–460, Oct. 2017, doi: 10.1016/j.jvcir.2017.01.019.

[63] C. Zhang, Y. Tian, X. Guo, and J. Liu, “DAAL: Deep activation-based attribute learning for action recognition in depth videos,” *Computer Vision and Image Understanding*, vol. 167, pp. 37–49, Feb. 2018, doi: 10.1016/j.cviu.2017.11.008.

[64] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, “Video-based human action recognition using deep learning,” p. 35.

[65] Q. Wu, Z. Wang, F. Deng, Z. Chi, and D. D. Feng, “Realistic Human Action Recognition With Multimodal Feature Selection and Fusion,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 4, pp. 875–885, Jul. 2013, doi: 10.1109/TSMCA.2012.2226575.

[66] Y. Yang, I. Saleemi, and M. Shah, “Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions,” *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 7, pp. 1635–1648, Jul. 2013, doi: 10.1109/TPAMI.2012.253.

[67] B. Ni, P. Moulin, X. Yang, and S. Yan, “Motion Part Regularization: Improving action recognition via trajectory group selection,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3698–3706, doi: 10.1109/CVPR.2015.7298993.

[68] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, “Structured Learning of Human Interactions in TV Shows,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2441–2453, Dec. 2012, doi: 10.1109/TPAMI.2012.24.

[69] K. N. Tran, A. Gala, I. A. Kakadiaris, and S. K. Shah, “Activity analysis in crowded environments using social cues for group discovery and human interaction modeling,” *Pattern Recognition Letters*, vol. 44, pp. 49–57, Jul. 2014, doi: 10.1016/j.patrec.2013.09.015.

[70] H. P. Martínez, G. N. Yannakakis, and J. Hallam, “Don’t Classify Ratings of Affect; Rank Them!,” *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, Jul. 2014, doi: 10.1109/TAFFC.2014.2352268.

[71] T. Lan, L. Sigal, and G. Mori, “Social roles in hierarchical models for human activity recognition,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 1354–1361, doi: 10.1109/CVPR.2012.6247821.
[72] B. Packer, K. Saenko, and D. Koller, “A combined pose, object, and feature model for action understanding,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 1378–1385, doi: 10.1109/CVPR.2012.6247824.
[73] B. Yao and L. Fei-Fei, “Modeling mutual context of object and human pose in human-object interaction activities,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun. 2010, pp. 17–24, doi: 10.1109/CVPR.2010.5540235.
[74] B. Elsner and B. Hommel, “Effect anticipation and action control,” *J Exp Psychol Hum Percept Perform*, vol. 27, no. 1, pp. 229–240, Feb. 2001, doi: 10.1037//0096-1523.27.1.229.
[75] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer, “Generation of fiducial marker dictionaries using Mixed Integer Linear Programming,” *Pattern Recognition*, vol. 51, pp. 481–491, Mar. 2016, doi: 10.1016/j.patcog.2015.09.023.
[76] G. Farnebäck, “Two-Frame Motion Estimation Based on Polynomial Expansion,” in *Image Analysis*, vol. 2749, J. Bigun and T. Gustavsson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 363–370.
[77] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” *arXiv:1804.02767 [cs]*, Apr. 2018, Accessed: May 10, 2019. [Online]. Available: http://arxiv.org/abs/1804.02767.
[78] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple Online and Realtime Tracking,” *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, Sep. 2016, doi: 10.1109/ICIP.2016.7533003.
[79] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *arXiv:1502.03167 [cs]*, Feb. 2015, Accessed: Aug. 13, 2019. [Online]. Available: http://arxiv.org/abs/1502.03167.
[80] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.
[81] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR*, p. 2012.
[82] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[83] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.
[84] R. Kazman, L. Bass, G. Abowd, and M. Webb, “SAAM: a method for analyzing the properties of software architectures,” in *Proceedings of 16th International Conference on Software Engineering*, May 1994, pp. 81–90, doi: 10.1109/ICSE.1994.296768.
[85] “Driver Monitoring System | Eye Tracking Technology,” *Smart Eye*. https://smarteye.se/automotive-solutions/ (accessed Mar. 06, 2020).

10 Appendix

10.1 Selected scenarios

10.1.1 Scenarios from business requirements

- Individual:
    - Activity: eating, drinking, sleeping, entering the car, exiting the car, car-passenger interaction, changing clothes, driving, day-dreaming, partying, watching films, reading, holding an object
    - 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
    - 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
    - Interaction: talking, fighting, playing, kissing, arguing, gestures, eye contact
    - Body language: approaching behavior
    - Inside-outside the car: gesture
- Safety: Unwanted behavior in a shared vehicle,
    - Destroying the interior,
    - Violent behavior,
    - Assault,
    - Molest

10.1.2 Scenarios from literature review

10.1.2.1 Comfort related

- Sleep
- Eat/Drink
- Using phone (talk, video call, dial, text, socializing)
- Watch movie
- Using computer (work, gaming, surf, socializing, online shopping)

- Play board games
- Rest/Relax
- Pay attention to the road/watch out
- Read a book
- Romance

10.1.2.2 Safety related

- Objects of detection interest
  - Laptop/Phone/Tablet
  - Hard/soft movable or latched objects
  - Food/drink
- Body position
  - Normal (recommended position by manufacturer)
  - Deviation
- Short terms activities
  - Phone and tablet use (talking, texting, surfing, dialing)
  - Reading non-interactive media (book, paper(s), map)
  - Writing (excluding texting)
  - Reaching an object
  - Eating/drinking
- Long terms activities
  - Hand(s) on/off steering wheel
  - Sleeping
  - Relaxing
- Emotions
  - Face expression of emotion
  - Aggression

10.2 Available algorithms and public datasets

Algorithm	Algo subcategory	Paper	Description	Dataset
Hierarchical model		A Hierarchical Deep Temporal Model for Group Activity Recognition	Three cues: Person level actions, temporal dynamics of a persons action and temporal evolution of group activity. Hierarchical LSTM approach Using AlexNet in several steps	Collective Activity Dataset

Algorithm	Algo subcategory	Paper	Description	Dataset
Multiview		Multi-view representation learning for multi-view action recognition.		IXMAS, M2I.
Handcrafted				
	Image feature			
	Temporal feature (optical flow, trajectory)	Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors		UCF101, HMDB51
	Combined features (3D filters)			
	Skeletal based			
Deep learning				
	CNN, RNN with LSTM	Structural Recurrent Neural Network (SRNN) for Group Activity Analysis	Use series of interconnected RNNs to jointly capture the actions of individuals, their interactions, as well as the group activity. An enhancement of the Hierarchical LSTM	
Combined	Fusion in different phase in the network	Going deeper into action recognition: A survey	Review of several combination ways of fusion the input information into the network	HMDB51 UCF101 UCF50 UCFSportsa Hollywood 2a Olympic Sportsa Sports-1 M

We also collect available public datasets that can be re-used and resulted in the below table.

DB Name	Link	Description	DB Size	Cate-gory	Annotated group activity keywords
ActivityNet 200	http://activity-net.org/download.html	Annotated activity database used in CVPR contests, no group activity annotations but can be useful in individual activity recognition. 200 activity classes 10,024 training videos (15,410 instances) 4,926 validation videos (7,654 instances) 5,044 testing videos (labels withheld)	N/A, only links to youtube	RGB	No group activity annotated
SBU Kinect Interaction Dataset	http://www3.cs.stonybrook.edu/~kyun/research/kinect_interaction/index.html	RGBD database (Kinect).	3.3GB	RGBD	approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands
K3HI	http://www.lmars.whu.edu.cn/prof_web/zhuxinyan/DataSetPublish/dataset.html	Small Kinect DB capture human interaction	7MB	RGBD	approaching, departing, pushing, kicking, punching, exchanging objects, pointing, and shaking hands
LIRIS human activities	https://projet.liris.cnrs.fr/voir/activities-dataset/download.html	Two sets of data by two different cameras: - MS Kinect module mounted on a remotely controlled Wany robotics Pekee II mobile robot which is part of the LIRIS-VOIR platform	12.5GB	RGB + RGBD	discussion, give object to other, handshaking, telephoning

DB Name	Link	Description	DB Size	Cate-gory	Annotated group activity keywords
		- Sony consumer camcorder			
ASLAN	https://www.openu.ac.il/home/hassner/data/ASLAN/ASLAN.html	3697 action samples from 1571 unique YouTube videos divided into 432 non-trivial action categories	N/A, only links to youtube	RGB	No keywords, only finding if the same actions occurred in two videos
Violent flows DB	https://www.openu.ac.il/home/hassner/data/violentflows/index.html	A database of real-world, video footage of crowd violence, along with standard benchmark protocols designed to test both violent/non-violent classification and violence outbreak detections. The data set contains 246 videos. All the videos were downloaded from YouTube. The shortest clip duration is 1.04 seconds, the longest clip is 6.52 seconds, and the average length of a video clip is 3.60 seconds.	N/A, only links to youtube	RGB	Violence on the crowd. Not yet tested for small group
CMU Motion Capture Database	http://mocap.cs.cmu.edu/faqs.php	There are 2605 trials in 6 categories and 23 subcategories. With group activity classes	4GB	RGB	shake hand, quarrel, pull, conversation

DB Name	Link	Description	DB Size	Cate-gory	Annotated group activity keywords
UT-Interaction dataset	http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html	The UT-Interaction dataset contains videos of continuous executions of 6 classes of human-human interactions: shake-hands, point, hug, push, kick and punch. Ground truth labels for these interactions are provided, including time intervals and bounding boxes. There is a total of 20 video sequences whose lengths are around 1 minute. Each video contains at least one execution per interaction, providing us 8 executions of human activities per video on average. Several participants with more than 15 different clothing conditions appear in the videos. The videos are taken with the resolution of 720*480, 30fps, and the height of a person in the video is about 200 pixels.	600MB	RGB	Hand Shaking Hugging Kicking Pointing Punching Pushing
HMDB51	http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#overview	6849 clips divided into 51 action categories, each containing a minimum of 101 clips	2GB	RGB	fencing, hug, kick someone, kiss, punch, shake hands, sword fight.
MSR Action 3D	https://www.uow.edu.au/~wanqing/#Datasets	MS dataset with D info. Also may have skeleton data MSRAction3DSkeletonREal3D.rar		RGBD	
ShakeFive	https://www.projects.science.uu.nl/shakefive/	A collection of Human interactions with accompanying skeleton		RGB	Fist bump, Hand shake, High five,

DB Name	Link	Description	DB Size	Cate-gory	Annotated group activity keywords
		metadata. 153 videos are encoded with ffmpeg x264 at a resolution of 1280x720.			Hug, Pass object, Thumbs up Rock-paper-scissors, Explaining
Human3.6M	http://vision.imar.ro/human3.6m/description.php	• 3.6 million 3D human poses and corresponding images • 11 professional actors (6 male, 5 female) • 17 scenarios (discussion, smoking, taking photo, talking on the phone...) • High-resolution 50Hz video from 4 calibrated cameras • Accurate 3D joint positions and joint angles from high-speed motion capture system • Pixel-level 24 body part labels for each configuration • Time-of-flight range data • 3D laser scans of the actors • Accurate background subtraction, person bounding boxes		RGBD	**Discussion,** smoking, **greeting,** eating, talking on phone
IXMAS	http://4drepository.inrialpes.fr/public/viewgroup/6	Downloadable files: original views of 5 cameras (23fps) in png-format (390x291), silhouettes in bpm-format (390x291), reconstructed volumes in matlab format (64x64x64), camera calibration data, framewise ground truth labeling: 0 - nothing, 1 - check watch, 2 - cross arms, 3 - scratch head, 4 -		Multi view	No group activity, but multiview activities

DB Name	Link	Description	DB Size	Cate-gory	Annotated group activity keywords
		sit down, 5 - get up, 6 - turn around, 7 - walk, 8 - wave, 9 - punch, 10 - kick, 11 - point, 12 - pick up, 13 - throw (over head), 14 - throw (from bottom up).			
i3DPost	http://kahlan.eps.surrey.ac.uk/i3dpost_action/	108 sequences, actions include: Walk, Run, Jump, Bend, Hand-wave, Jump-in-place, Sit-StandUp, Run-fall, Walk-sit, Run-jump-walk, Handshake, Pull, Facial-expressions		Multi view	Handshake, (multiview activities)

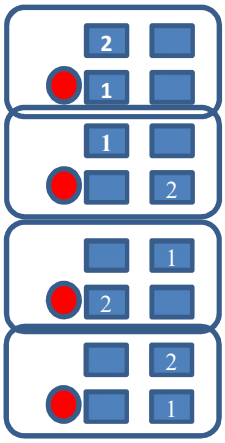
10.3 Data capture setup and optimized list of scenarios

Setup 1

Setup 2

Setup 3

Setup 4



		Setup number			
Scenario #	Scenario name	1	2	3	4
1	Driver and front-seat passenger conversing	28			
2	Driver and front-seat passenger shake hands	29			
3	Driver and front-seat passenger do a high-five	30			
4	Front-seat passenger handing a phone to the driver	31			
5	Front-seat passenger shows the contents on a phone screen to the driver	32			

		Setup number			
Scenario #	**Scenario name**	**1**	**2**	**3**	**4**
6	Front-seat passenger handing a bottle to the driver	33			
7	Driver and back-seat passenger conversing			28	
8	Driver and back-seat passenger shake hands			29	
9	Driver and back-seat passenger do a high-five			30	
10	Back-seat passenger handing a phone to the driver			31	
11	Back-seat passenger shows the contents on a phone screen to the driver			32	
12	Back-seat passenger handing a bottle to the driver			33	
13	Driver and either passenger doing arbitrary other actions				
14	Two back-seat passengers playing card games				14
15	Two back-seat passengers playing tablet games				15
16	Two back-seat passengers conversing				16
17	Driver sleeping/relaxing	1		1	
18	Driver hands on steering wheel	3		3	
19	Driver hands off steering wheel	5		5	
20	Driver reaching and grabbing a phone	7		7	
21	Driver holding a phone	9		9	
22	Driver talking on phone	11		11	
23	Driver using phone	13		13	
24	Driver talking on phone with headset or handsfree	15		15	
25	Driver holding a bottle	17		17	
26	Driver drinking from a bottle	19		19	
27	Driver holding a candy bar	21		21	
28	Driver eating a candy bar	23		23	
29	Driver reaching and grabbing a book	25		25	
30	Driver reaching and grabbing a newspaper	27		27	
31	Front-seat passenger sleeping/relaxing	2	1		
32	Front-seat passenger reaching and grabbing a laptop	4	3		
33	Front-seat passenger using on laptop	6	5		
34	Front-seat passenger holding a phone	8	7		
35	Front-seat passenger talking on phone	10	9		
36	Front-seat passenger using phone	12	11		
37	Front-seat passenger talking on phone with headset or handsfree	14	13		
38	Front-seat passenger holding a bottle	16	15		

		Setup number			
Scenario #	**Scenario name**	**1**	**2**	**3**	**4**
39	Front-seat passenger drinking from a bottle	18	17		
40	Front-seat passenger holding a candy bar	20	19		
41	Front-seat passenger eating a candy bar	22	21		
42	Front-seat passenger reading a book	24	23		
43	Front-seat passenger reading a newspaper	26	25		
44	Back-seat passenger sleeping/relaxing		2	2	
45	Back-seat passenger reaching and grabbing a laptop		4	4	
46	Back-seat passenger using on laptop		6	6	
47	Back-seat passenger holding a phone		8	8	
48	Back-seat passenger talking on phone		10	10	
49	Back-seat passenger using phone		12	12	
50	Back-seat passenger talking on phone with headset or handsfree		14	14	
51	Back-seat passenger holding a bottle		16	16	
52	Back-seat passenger drinking from a bottle		18	18	
53	Back-seat passenger holding a candy bar		20	20	
54	Back-seat passenger eating a candy bar		22	22	
55	Back-seat passenger reading a book		24	24	
56	Back-seat passenger reading a newspaper		26	26	

		object								activity									
Basic scenario #	Description	apple	banana	phone	bottle	cup	laptop	book	steering wheel	interaction	handing	showing	eating	drinking	holding	talking	using	reading	reaching and grabbing
1	Conversation									x						x			
2	Hand shake									x									
3	High-five									x									
4	Holding a phone			x											x				
5	Talking on a phone			x												x			
6	Using a phone			x													x		

		object								activity									
Basic scenario #	Description	apple	banana	phone	bottle	cup	laptop	book	steering wheel	interaction	handing	showing	eating	drinking	holding	talking	using	reading	reaching and grabbing
7	Reaching and grabbing a phone			x															x
8	Handing a phone			x						x	x								
9	Showing contents on phone screen			x						x		x							
10	Handing a bottle				x						x								
11	Drinking from a bottle				x									x					
12	Holding a banana		x												x				
13	Eating a banana		x										x						
14	Holding an apple	x													x				
15	showing an apple	x										x							
16	Eating an apple	x											x						
17	Handing a cup					x					x								
18	Drinking from a cup					x								x					
19	Arbitrary other actions / silently looking around																		
20	Sleeping/relaxing																		
21	Reaching and grabbing a book							x											x
22	Reading a book							x										x	
23	Reaching and grabbing a laptop						x												x
24	Using laptop						x										x		

		object								activity									
Basic scenario #	Description	apple	banana	phone	bottle	cup	laptop	book	steering wheel	interaction	handing	showing	eating	drinking	holding	talking	using	reading	reaching and grabbing
25	Reading on laptop						x											x	
26	Hands on steering wheel								x										
27	Hands off steering wheel								x										