AVENUE

Autonomous Vehicles to Evolve to a New Urban Experience

DELIVERABLE

D4.6 Final iteration In-vehicle services

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769033

Disclaimer

Document Information

Grant Agreement Number	769033
Full Title	Autonomous Vehicles to Evolve to a New Urban Experience
Acronym	AVENUE
Deliverable	D4.6 Final iteration in-vehicle services
Due Date	01.08.2022
Work Package	WP4
Lead Partner	Autonomous Mobility
Leading Author	Christian Zinckernagel
Dissemination Level	Public, restricted

Document History

Version	Date	Author	Description of change
0.1	03.03.2022	Christian Zinckernagel, Nanna Felthaus, Christian Bering, Sophie Green, Amobility	First draft (Starting point D4.5)
0.2	15.04.2022	Dimitrios Tsiktsiris, Anastasios Vafeiadis, Ioannis Papadimitriou, Evangelos Athanasakis, Eleni Diamantidou, Theoktisti Marinopoulou, Athanasios	First revision of services

		Papadakis, Nikolaos Dimitriou, Antonios Lalas, Konstantinos Votis, Dimitrios Tzovaras, CERTH	
0.3	20.05.2022	Christian Zinckernagel, Nanna Felthaus, Christian Bering, Sophie Green, Amobility	Installation guide
	15.07.2022	Dimitrios Tsiktsiris, Anastasios Vafeiadis, Ioannis Papadimitriou, Evangelos Athanasakis, Eleni Diamantidou, Theoktisti Marinopoulou, Athanasios Papadakis, Nikolaos Dimitriou, Antonios Lalas, Konstantinos Votis, Dimitrios Tzovaras, CERTH	Second revision of services, API integration
0.4	27.09.2022	Dimitrios Tsiktsiris, Anastasios Vafeiadis, Ioannis Papadimitriou, Evangelos Athanasakis, Eleni Diamantidou, Theoktisti Marinopoulou, Antonios Lalas, Konstantinos Votis, Dimitrios Tzovaras, CERTH	Final revision of services, Validation
0.5	04.10.2022	Christian Zinckernagel, Nanna Felthaus, Christian Bering, Sophie Green, Amobility	Version ready for review
1.0	31.10.2022	Christian Zinckernagel, Amobility	Final version

Table of Contents

List of figures

List of tables

Acronyms

ADS	Automated Driving Systems
AI	Artificial Intelligence
AM	Automated Mobility
API	Application Protocol Interface
AV	Automated Vehicle
BMM	Business Modelling Manager
CAV	Connected and Automated Vehicles
CB	Consortium Body
CERN	European Organization for Nuclear Research
D7.1	Deliverable 7.1
DC	Demonstration Coordinator
DI	The department of infrastructure (Swiss Canton of Geneva)
DMP	Data Management Plan
DSES	Department of Security and Economy - Traffic Police (Swiss Canton of Geneva)
DTU test track	Technical University of Denmark test track
EAB	External Advisory Board
EC	European Commission
ECSEL	Electronic Components and Systems for European Leadership
EM	Exploitation Manager
EU	European Union
EUCAD	European Conference on Connected and Automated Driving
F2F	Face to face meeting
FEDRO	(Swiss) Federal Roads Office
FOT	(Swiss) Federal Office of Transport
GDPR	General Data Protection Regulation
GIMS	Geneva International Motor Show
GNSS	Global Navigation Satellite System
HARA	Hazard Analysis and Risk Assessment
IPR	Intellectual Property Rights
IT	Information Technology
ITU	International Telecommunications Union
LA	Leading Author
LIDAR	Light Detection And Ranging
MEM	Monitoring and Evaluation Manager
MT	MobileThinking
OCT	General Transport Directorate of the Canton of Geneva
ODD	Operational Domain Design
OEDR	Object And Event Detection And Response
OFCOM	(Swiss) Federal Office of Communications
PC	Project Coordinator
PEB	Project Executive Board
PGA	Project General Assembly
PRM	Persons with Reduced Mobility
PSA	Group PSA (PSA Peugeot Citroën)
PTO	Public Transportation Operator
PTS	Public Transportation Services
QRM	Quality and Risk Manager
QRMB	Quality and Risk Management Board
RN	Risk Number
SA	Scientific Advisor
SAE Level	Society of Automotive Engineers Level (Vehicle Autonomy Level)
SAN	(Swiss) Cantonal Vehicle Service
SDK	Software Development Kit
SLA	Sales Lentz Autocars
SMB	Site Management Board
SoA	State of the Art
SOTIF	Safety Of The Intended Functionality
SWOT	Strengths, Weaknesses, Opportunities, and Threats.
T7.1	Task 7.1
TM	Technical Manager
TPG	Transport Publics Genevois
UITP	Union Internationale des Transports Publics (International Transport Union)
V2I	Vehicle to Infrastructure communication
WP	Work Package
WPL	Work Package Leader

Executive Summary

This deliverable will introduce the preparation work, done by Amobility (AM) and CERTH, for prototyping, testing and validating 5 in-vehicle services in the AVENUE project. For each chosen service the technological framework is described, including the maturity of the technology and the equipment needed to prototype the service elements. Furthermore a prototyping plan will be introduced for each service and the result framework will be defined.

The five chosen services that will be introduced and tested in the AVENUE project:

- Security trust services (example: Prevention of night aggressions)
- Automated passenger presence
- Follow your kid/grandparents
- Shuttle environment assessment
- Smart feedback system

As a part of the AVENUE goals are to be able to drive fully autonomous without the presence of a safety officer, the role of the safety officer is described. The chosen services all target some of the given services that the safety officer offers in person. The five services are all services that need to be automated in order to be able to remove the operator.

Furthermore the analysis and data processing is based on machine learning - in this case training and development of the data points used to assess and determine when certain events occur. The behaviour of Danish travellers has been implemented via multiple data recording sessions with employees and test personnel in test environments to train the algorithms to detect the right movements, sounds and behaviour.

The report presents at the end a validation of the 5 services taking the quality, stability and trustworthiness of the services into account.

1 Introduction

AVENUE aims to design and carry out full-scale demonstrations of urban transport automation by deploying, for the first time worldwide, fleets of Automated minibuses in low to medium demand areas of 4 European demonstrator cities (Geneva, Lyon, Copenhagen and Luxembourg) and 2 to 3 replicator cities. The AVENUE vision for future public transport in urban and suburban areas, is that Automated vehicles will ensure safe, rapid, economic, sustainable and personalised transport of passengers. AVENUE introduces disruptive public transportation paradigms on the basis of on-demand, door-to-door services, aiming to set up a new model of public transportation, by revisiting the offered public transportation services, and aiming to suppress prescheduled fixed bus itineraries.

Vehicle services that substantially enhance the passenger experience as well as the overall quality and value of the service will be introduced, also targeting elderly people, people with disabilities and vulnerable users. Road behaviour, security of the Automated vehicles and passengers' safety are central points of the AVENUE project.

At the end of the AVENUE project four-year period the mission is to have demonstrated that Automated vehicles will become the future solution for public transport. The AVENUE project will demonstrate the economic, environmental and social potential of Automated vehicles for both companies and public commuters while assessing the vehicle road behaviour safety.

1.1 On-demand Mobility

Public transportation is a key element of a region's economic development and the quality of life of its citizens.
Governments around the world are defining strategies for the development of efficient public transport based on different criteria of importance to their regions, such as topography, citizens' needs, social and economic barriers, environmental concerns and historical development. However, new technologies, modes of transport and services are appearing, which seem very promising to the support of regional strategies for the development of public transport.
On-demand transport is a public transport service that only works when a reservation has been recorded and will be a relevant solution where the demand for transport is diffuse and regular transport is inefficient.
On-demand transport differs from other public transport services in that vehicles do not follow a fixed route and do not use a predefined timetable. Unlike taxis, on-demand public transport is usually also not individual. An operator or an automated system takes care of the booking, planning and organization.
It is recognized that the use and integration of on-demand Automated vehicles has the potential to significantly improve services and provide solutions to many of the problems encountered today in the development of sustainable and efficient public transport.

1.2 Fully Automated Vehicles

A self-driving car, referred in the AVENUE project as **a Fully Automated Vehicle** (**AV**), also referred as Autonomous Vehicle, is a vehicle that is capable of sensing its environment and moving safely with no human input.

The terms *automated vehicles* and *autonomous vehicles* are often used together. The Regulation 2019/2144 of the European Parliament and of the Council of 27 November 2019 on type-approval requirements for motor vehicles defines "automated vehicle" and "fully automated vehicle" based on their autonomous capacity:

- An "automated vehicle" means a motor vehicle designed and constructed to move autonomously for certain periods of time without continuous driver supervision but in respect of which driver intervention is still expected or required
- "fully automated vehicle" means a motor vehicle that has been designed and constructed to move autonomously without any driver supervision

In AVENUE we operate ***Fully Automated minibuses for public transport***, (previously referred as Autonomous shuttles, or Autonomous buses), and we refer to them as simply *Automated minibuses* or *the AVENUE minibuses.*

In relation to the SAE levels, the AVENUE project will operate SAE Level 4 vehicles.

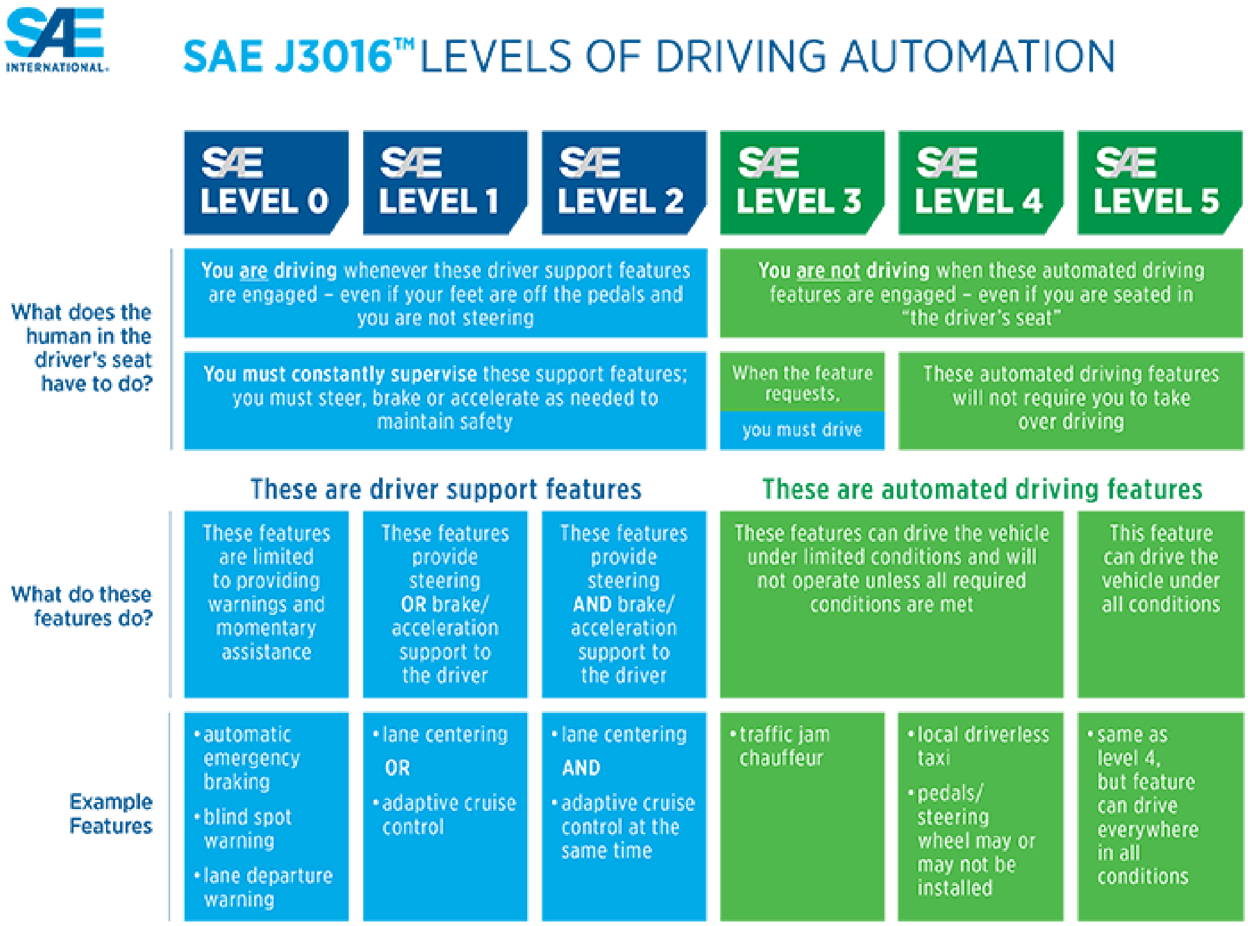


Figure 1. SAE Driving Automation levels (©2020 SAE International)

1.2.1 Autonomous vehicle operation overview

We distinguish in AVENUE two levels of control of the AV: micro-navigation and macro-navigation. Micro navigation is fully integrated in the vehicle and implements the road behaviour of the vehicle, while macro-navigation is controlled by the operator running the vehicle and defines the destination and path of the vehicle, as defined the higher view of the overall fleet management.
For micro-navigation Automated Vehicles combine a variety of sensors to perceive their surroundings, such as 3D video, LIDAR , sonar, GNSS, odometry and other types sensors. Control software and systems, integrated in the vehicle, fusion and interpret the sensor information to identify the current position of the vehicle, detecting obstacles in the surround environment, and choosing the most appropriate reaction of the vehicle, ranging from stopping to bypassing the obstacle, reducing its speed, making a turn etc.
For the Macro-navigation, that is the destination to reach, the Automated Vehicle receives the information from either the in-vehicle operator (in the current configuration with a fixed path route), or from the remote control service via a dedicated 4/5G communication channel, for a fleet-managed operation. The fleet management system takes into account all available vehicles in the services area, the passenger request, the operator policies, the street conditions (closed streets) and send route and stop information to the vehicle (route to follow and destination to reach).

1.2.2 Automated vehicle capabilities in AVENUE

The Automated vehicles employed in AVENUE fully and automatically manage the above defined, micro-navigation and road behaviour, in an open street environment. The vehicles are Automatically capable to recognise obstacles (and identify some of them), identify moving and stationary objects, and Automatically decide to bypass them or wait behind them, based on the defined policies. For example with small changes in its route the AVENUE mini-bus is able to bypass a parked car, while it will slow down and follow behind a slowly moving car. The AVENUE mini-buses are able to handle different complex road situations, like entering and exiting round-about in the presence of other fast running cars, stop in zebra crossings, communicate with infrastructure via V2I interfaces (ex. red light control).
The mini-buses used in the AVENUE project technically can achieve speeds of more than 60Km/h. However this speed cannot be used in the project demonstrators for several reasons, ranging from regulatory to safety. Under current regulations the maximum authorised speed is 25 or 30 Km/h (depending on the site). In the current demonstrators the speed does not exceed 23 Km/h, with an operational speed of 14 to 18 Km/h. Another, more important reason for limiting the vehicle speed is safety for passengers and pedestrians. Due to the fact that the current LIDAR has a range of 100m and the obstacle identification is done for objects no further than 40 meters, and considering that the vehicle must safely stop in case of an obstacle on the road (which will be “seen” at less than 40 meters distance) we cannot guarantee a safe braking if the speed is more than 25 Km/h. Note that technically the vehicle can make harsh break and stop with 40 meters in high speeds (40 -50 Km/h) but then the break would too harsh putting in risk the vehicle passengers. The project is working in finding an optimal point between passenger and pedestrian safety.
Due to legal requirements a **Safety Operator** must always be present in the vehicle, able to take control any moment. Additionally, at the control room, a **Supervisor** is present controlling the fleet operations. An **Intervention Team** is present in the deployment area ready to intervene in case of incident to any of the minibuses. Table 1 provides an overview of the AVENUE sites and OOD

	Summary of AVENUE operating sites demonstrators							
	TPG		Holo			Keolis	Sales-Lentz	
	Geneva		Copenhagen	Oslo	Copenhagen	Lyon	Luxembourg	
Site	Meyrin	Belle-Idée	Nordhavn	Ormøya	Slagelse	ParcOL	Pfaffental	Contern
Funding	TPG	EU + TPG	EU + Holo	EU + Holo	EU + Holo	EU + Keolis	EU + SLA	EU + SLA
Start date of project	August 2017	May 2018	May 2017	August 2019	August 2020	May 2017	June 2018	June 2018
Start date of trial	July 2018	June 2020	September 2020	December 2019	August 2021	November 2019	September 2018	September 2018
Type of route	Fixed circular line	Area	Fixed circular line	Fixed circular line	Area	Fixed circular line	Fixed circular line	Fixed circular line
Level of on-demand service*	Fixed route / Fixed stops	Flexible route / On-demand stops	Fixed route / Fixed stops	Fixed route / Fixed stops	Flexible route / On-demand stops	Fixed route/Fixed stops	Fixed route / Fixed stops	Fixed route / Fixed stops
Route length	2,1 km	38 hectares	1,3 km	1,6 km	5,5 km	1,3 km	1,2 km	2,3 km
Road environment	Open road	Semi-private	Open road	Open road	Open road	Open road	Public road	Public road
Type of traffic	Mixed	Mixed	Mixed	Mixed	Mixed	Mixed	Mixed	Mixed
Speed limit	30 km/h	30 km/h	30 km/h	30 km/h	30 km/h	8 to 10 km/h	30 km/h	50 km/h
Roundabouts	Yes	Yes	No	No	No	Yes	No	No
Traffic lights	No	No	No	No	No	Yes	Yes	Yes
Type of service	Fixed line	On demand	Fixed line	Fixed line	On demand	Fixed line	Fixed line	Fixed line
Concession	Line (circular)	Area	Line (circular)	Line (circular)	Area	Line (circular)	Line (circular)	Line (circular)
Number of stops	4	> 35	6	6	7	2	4	2
Type of bus stop	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed

Bus stop infrastructure	Yes	Sometimes, mostly not	Yes	Yes	Yes	Yes	Yes	Yes
Number of vehicles	1	3-4	1	2	2	2	2	1
Timetable	Fixed	On demand	Fixed	Fixed	On-demand	Fixed	Fixed	Fixed
Operation hours	Monday-Friday (5 days)	Sunday-Saturday (7 days)	Monday-Friday (5 days)	Monday-Sunday (7 days)	Monday-Friday (5 days)	Monday-Saturday (6 days)	Tuesday & Thursday Saturday, Sunday & every public holiday	Monday - Friday
Timeframe weekdays	06:30 – 08:30 / 16:00 – 18:15	07:00 – 19:00	10:00 – 18:00	7:30 – 21:30	07:00-16:00	08:30 – 19:30	12:00 – 20h00	7:00 – 9:00 16:00 – 19:00
Timeframe weekends	No service	07:00 – 19:00	No service	9:00 – 18:00	No service	08:30 – 19:30	10:00 – 21:00	No Service
Depot	400 meters distance	On site	800 meters distance	200 meters distance	On site	On site	On site	On site
Driverless service	No	No	No	No	No	No	No	No
Drive area type/ODD	B-Roads	Minor roads/parking	B-Roads/minor roads	B-Roads	B-Roads/parking	B-Roads	B-Roads	B-Roads/parking
Drive area geo/ODD	Straight lines/plane	Straight lines/ plane	Straight lines/ plane	Curves/slopes	Straight lines / curves	Straight Lines/ plane	Straight lines/ plane	Straight lines/ plane
Lane specification/ODD	Traffic lane	Traffic lane	Traffic lane	Traffic lane	Traffic lane	Traffic lane	Traffic lane	Traffic lane
Drive area signs/ODD	Regulatory	Regulatory	Regulatory, Warning	Regulatory	Regulatory	Regulatory	Regulatory	Regulatory
Drive area surface/ODD	Standard surface, Speedbumps	Standard surface, Speedbumps	Standard surface Speedbumps, Roadworks	Frequent Ice, Snow	Standard surface, snow during winter	Standard surface, Potholes	Standard surface	Standard surface

Table 1: Summary of AVENUE operating site (+ODD components)

1.3 Preamble

WP4: Development, Adaptation and integration of Passenger Transport and in-, out-of-, vehicle services, aims to design, develop, adapt and integrate services to support users of autonomous vehicles before the trip, during the trip, and at the end of the trip. The main objective of WP4 is to provide services in order to demonstrate that the user experience can be seamless and secure, and that people embrace this new technology. Hence, we have to include the following services:

- Adapt and integrate existing transport services
- Develop autonomous vehicle specific services
- Provide services that foster the acceptance of driverless vehicles by both passengers and people interacting with the shuttles
- Introduce safety related services

The target of task T4.2 is to develop, teste and integrate innovative in-vehicle services, in collaboration with the four operators in Lyon, Luxembourg, Geneva and Copenhagen, respectively and with technical providers of the AVENUE consortium. The in-vehicle services should in combination with the out-of-vehicle services support a holistic service for travellers commuting with the AVs.

In-vehicle services are services developed to improve the user experience when travelling with autonomous vehicles. The services are user-centric and focus on supporting travellers (including PRM) with smart solutions while sitting inside the vehicle - in this case the Navya autonomous vehicle.

This deliverable, D4.6: Final iteration in-vehicle services, aims at introducing the preoperational work conducted with the purpose of prototyping and testing innovative and state-of-the-art technologies in autonomous shuttles.

This deliverable introduced the preparation of five selected in-vehicle services that was prototyped and tested. All five services are based on camera and sensor technologies, developed by CERTH. The services will be tested mainly on the two Amoblity routes (Copenhagen and Oslo).

2 In-vehicle services

The end user in-vehicle services from the proposal can be seen in the following including type and timeframe:

1. Intelligent ticket control (In-vehicle service)
2. In-vehicle entertainment (In-vehicle service)
3. Virtual personality interaction (In-vehicle service)
4. Emergency automatic call system (In-vehicle service)
5. Enhance the sense of security and trust (In-vehicle service)
6. Prevention of night aggressions (In-vehicle service)
7. Visualization in real time of the path / destination (In-and-out-of-vehicle service)
8. Follow my kid / grandmother (In-and-Out-of-vehicle service)
9. Mutual help facilitation (In-and-Out-of-vehicle service)

10. Passenger presence (In-vehicle service + AV functionality service)

2.1 Existing in-vehicle technical facilities

As of today, the Navya autonomous vehicle is equipped with some functions which will be integrated in the provision of services, targeted at the passengers. Therefore, the functions are highlighted with the purpose of including them in the further service development process.

The in-vehicle technical functions are as follows:

<u>Open/close doors button</u>
Manual buttons inside the vehicle, allowing the passengers to press them. The doors do also open automatically at each stop.

<u>Wheelchair ramp button</u>
Manual button inside the vehicle. Passenger or safety officer can use the button to activate the automatic wheelchair ramp. This function is not installed in all Navya autonomous vehicles as a standard. It must be installed on the shuttle by Navya.

<u>Real time visualisation of route and stops (service screen)</u>
There is a screen inside the vehicle showing a map of the route, including the stops. The screen is also used by the safety officer to restart the vehicle or detect issues during operation. The screen is currently not designed for passengers, but for the safety officer. In the future, once the safety officer is removed from the vehicles, the screen will function as the main interaction point between the travellers and the service.

<u>Speaker system</u>

Speakers located inside the vehicle, enabling contact from Navya supervision to passengers during emergencies. Potentially in the future, the PTO's could use the system for service announcements, safety protocols or regular communication - especially once the safety officer is removed from the autonomous vehicles.

Emergency call button
A manual button located under the service screen allows the passengers to get in contact with Navya supervision. The button could potentially in the future be used to link the travellers with the operators' own supervision - leaving the contact to Navya to the operators instead.

Emergency stop button
Manual button located on each side of the large window in the middle. Manually brakes the vehicle.

Emergency door opening handle
Manual handle located on each side of the doors. Manually forces the doors to open.

Emergency Glass breaking hammer
A hammer clearly located inside the vehicle is available to be used in case passengers are blocked and need to exit from the window of the e-minibus

Emergency First Aid Kit
A First Aid kit is located inside the e-minibus and is easily accessible in case of need.

Fire extinguisher
Inside the e-minibus a fire extinguisher is installed for use in case of any emergency.

Interior camera
Inside the e-minibus there is a camera that records the inside of the e-minibus. The camera is a fisheye camera and can be used by the operators to assess the situation inside the shuttle.

For the prototyping and testing of the chosen five services, the speaker system, the emergency call button and the interior camera will be used for different tests.

2.2 AVENUE goal

As a part of the AVENUE goals and vision, the safety officer has to be taken out of the vehicles when possible. In order to do so it is essential to understand the role of the safety officer in shuttles. What kind of services are he/she performing, how is the existence of a safety officer perceived by the travellers and so forth. As a part of understanding this, the role of the safety officer is shortly described (mostly from D4.4 including some new insights) - as T4.2 focusses on in-vehicle services, the following description also focuses on the role of the safety officer inside the shuttle.

2.2.1 Safety officer

This presence of the safety officer enables the operators to gather valuable insight and observation about the users' behaviour and interaction with in-vehicle services, and thereby gather a solid foundation for further development and refinement of existing and new in-vehicle services. The safety officer in-vehicle services are identified with the purpose of understanding what is necessary to automate in order to take the safety officer out of the vehicle. Furthermore, by identifying the role of the safety officer it is possible to pair up the proposal services.

The analysis of the in-vehicle services provided by the safety officer is mainly based on the initial insights from the State of the Art analysis and the experience of Autonomous Mobility. Once all operators are in operation, the safety officer inside the vehicles will be further analysed and compared across the operation sites - contributing to a more refined and exhaustive definition of the in-vehicle services that needs to be designed in the future.

The following table 2 shows the current in-vehicle services provided by the safety officer. Each service is shortly described, with the purpose of initiating a thorough investigation of what is necessary once the safety drivofficerer is removed from the vehicle. For each part one or more of the in-vehicle proposal services are paired, where they make sense.

Table 2. Safety officer (in-vehicle services)

Safety officer Current in-vehicle services provided by the safety officer			
#	In-vehicle service	Description	Proposal services that potentially can automate the service
1	Operational information	Providing information about the project, route, vehicle, laws etc.	- Visualization in real time of the path / destination (Mobile application)
2	Safety + risk mitigation	Safety officer presence and authority establishes a safe environment for the passengers. + Active risk control (joystick, eyes etc.)	- Security and trust services - Prevention of night aggressions - Emergency automatic call system - Mutual help facilitation
3	Entertainment	The safety officer can interact with the passengers and start conversations about anything + tell stories about the technology etc.	- In-vehicle entertainment - Virtual personality interaction
4	Travel assistance	The safety officer can assist passengers in boarding and exiting the vehicle + ensure that the bus stops where the passengers want to get on/off.	- Intelligent ticket control - Follow my kid / grandmother - Passenger presence
5	Area guidance	The safety officer can assist passengers in finding restaurants, shops etc. in the local area of the route.	- Visualization in real time of the path / destination (Mobile application) - Virtual personality interaction
6	Branding	Branding of operating company, logo, name, etc. Can be used to assure users that the operation is safe and offered by a trusted company. Things are under control.	- Visualization in real time of the path / destination (Mobile application)
7	Provides data on various areas	User behaviour/insights + counting passenger + road user behaviour	- Visualization in real time of the path / destination (Mobile application) - Virtual personality interaction

2.3 In-vehicle service focus

The services that will be focused on in the next 8-12 months are as follows:

- Security trust services (example: Prevention of night aggressions)
- Automated passenger presence
- Follow your kid/grandparents
- Shuttle environment assessment
- Smart feedback system

The reason for choosing these services are two-fold:

- They are all essential services necessary to remove the safety officer from the shuttle, as safety and efficient operation are critical factors. For example ensuring safety for the travellers have been identified in WP2 as the single most important factor for choosing to ride with the autonomous shuttle.
- They are all based on the same foundation of technology with cameras, sensors and algorithms, meaning that we can harness from testing multiple types of services, once we have installed the necessary equipment. The analysis is also similar and therefore the results can be compiled together. The technology, equipment and algorithms are provided by CERTH and tested and prototyped together with Amobility.

During the next 8-12 months the above described services will be tested, further prototyped and analysed on the Oslo and Copenhagen site of Amobility. The purpose of the development of services is to prototype and test the services as a concept. The services will probably not be fully developed and implemented, but recommendations for further improvements and development will be the results of the service development and testing. For the development and testing a close collaboration will be established with MobileThinking as they have developed the mobile application that can be connected to the in-vehicle services over time.

3 Service: Enhance the sense of security and trust

3.1 Concept of service

The service "Enhance the sense of security and trust" aims to address the new reality that is formed in autonomous shuttles mobility infrastructures as a result of the absence of the bus driver and the increased threat from terrorism in European cities. Typically, drivers are trained to handle incidents of passengers' abnormal behaviour, incidents of petty crimes, etc. according to standard procedures adopted by the transport operator. Surveillance using sensors such as cameras (cameras of different technologies can be used so that passengers' privacy is protected) and microphones, as well as smart software in the bus will maximize the feeling of security and the actual level of security.

Several concerns of the end users regarding the Safety and Robustness of the autonomous vehicles that are directly linked to the final User Acceptance of the new technology, can be identified. The prospective passengers fear several possible instances that could arise in case there is no driver in the bus. Indicatively:

- No one will be in the bus to perform first aid if required
- Feeling of discomfort being all alone in the bus at night, especially in certain neighbourhoods
- No authority figure present to keep passengers calm (eg. school kids)
- Vandalism, bag snatching, indoor fighting, unaccompanied luggage

To address the aforementioned concerns on social and personal safety and security into the vehicle, certain measures need to be implemented. For example the detection of unaccompanied luggage and of other personal belongings may raise a notification or an alert to the supervisor and/or the suitable authorities. This may be followed by appropriate notifications and/or instructions to the passengers, while the vehicle may also implement respective actions.

Moreover, implementing a solution for enhancing the safety and security inside the autonomous buses will support safekeeping not only the users of the autonomous public bus but also the vehicle itself. In this section, the implementation of a video, depth and audio analytics software module for an embedded security subsystem or for cloud-based services of the system are described along with appropriate planning for the deployment and test of the service into the pilot sites of the Avenue project.

3.1.1 Use case

The service addresses the timely, accurate, robust and automatic detection of various petty crime types or misdemeanors as well as the assistance of authorized end-users towards the re-identification of any offenders. A misdemeanor is any "lesser" criminal act in some common law legal systems. Misdemeanors are generally punished less severely than felonies, but theoretically more so than administrative infractions (also known as minor, petty, or summary offences) and regulatory offences. Many misdemeanors are punished with monetary fines.

The petty crimes that are targeted for identification by the sensors include: petty theft like bag snatching and pickpocketing, vandalism, aggression, illegal consumption of cigarettes, public intoxication, simple assault and disorderly conduct. These are explained in more detail:

Petty theft: In common usage, theft is the taking of another person's property or services without that person's permission or consent with the intent to deprive the rightful owner of it.
Vandalism: Vandalism is the action involving deliberate destruction of or damage to public or private property.
Aggression: Aggression is overt or covert, often harmful, social interaction with the intention of inflicting damage or other unpleasantness upon another individual.
Public intoxication: Public intoxication, also known as "drunk and disorderly" and drunk in public, is a summary offense in some countries rated to public cases or displays of drunkenness.
Simple assault: An assault is the act of inflicting physical harm or unwanted physical contact upon a person or, in some specific legal definitions, a threat or attempt to commit such an action.

Disorderly conduct: Disorderly conduct makes it a crime to be drunk in public, to "disturb the peace", or to loiter in certain areas.

In the context of Avenue project, the following use cases have been identified to be further examined and addressed:

Use Case 1: Unaccompanied Luggage Monitoring

- Inside the autonomous shuttle there is unaccompanied luggage which remains unmoved for a long time.
- The video cameras installed in the autonomous shuttle acquire the color depth images and the data are fed into the system's video analytics algorithms for further analysis.
- In case the algorithms identify that the total time of the luggage that remains unmoved in the vehicle passes the predefined time frame, a notification is sent to the security operator.
- The security operator monitors the clips that are captured and evaluates the criticality of the situation and whether to intervene or not.

Use Case 2: Bag Snatch Detection

- While a commuter is in the autonomous shuttle a petty crime takes place in the form of assault.
- The commuter is attacked by another person who is attempting to snatch the bag from the commuter.
- The aggressive incident is captured by the sensors in the bus, both microphone and cameras, and fed to the video analytics component for further analysis.
- The system sends a security alert to the operator or security supervisor.
- The course of action of the operator is a human decision, that means, whether he/she will decide to stop the autonomous shuttle or will notify the passengers via the radio system.

Use Case 3: Vandalism Detection

- A young person onboards the autonomous shuttle during an itinerary performed by the vehicle during the night hours.
- The youngster attempts to perform a vandalism action on the shuttle, through painting graffiti on the windows.
- The night mode of the cameras installed in the vehicle acquire the data that will be fed to the video analytics algorithms for further analysis.
- The person is warned by the radio system of the autonomous shuttle or security personnel intervenes by stopping the bus.

Use Case 4: Sound Events Detection

- A person onboards the autonomous shuttle during an itinerary performed by the vehicle.
- The person attempts to terror the passengers on the shuttle, through breaking the windows. The passengers may also scream during this behavior or call for help.
- The microphones installed in the vehicle acquire the data that will be fed to the audio analytics algorithms for further analysis.
- The person is warned by the radio system of the autonomous shuttle or security personnel intervenes by stopping the bus.

3.2 Stakeholders (development/prototyping team)

In this section, the relevant stakeholders involved in the development and prototyping of the Enhance the Sense of Security and Trust Service are presented. Specifically:

- **CERTH** develops software and algorithms for the detection of abnormal events using video and sound analysis techniques. The service is implemented by adopting a platform for petty crimes and incident detection systems developed for the active detection of abnormal behaviour, as well as suspicious objects in conjunction with a variety of sensors. Sensors such as cameras and microphones, along with machine learning algorithms are employed for the timely, accurate and robust detection of petty crimes and incidents. CERTH is also involved in the installation of the related sensors and software into the autonomous shuttles on the operators facilities.
- **Bestmile** provides multiple integration interfaces towards the different stakeholders in the project (e.g., vehicle manufacturers, public transport operators) into its cloud platform. In this service, Bestmile provides connectivity between the vehicle and the related operators regarding the notifications that are generated by the detection software.
- **MobileThinking** develops the AVENUE mobile app for the end users that will be involved in the project. In this service, MobileThinking develops the interconnection with the AVENUE mobile app by providing the notifications generated from the system to the end users. MobileThinking also develops a web-based monitoring dashboard for PTO’s. The dashboard collects and stores the information collected by in-vehicle sensors and notifies the intervention team in cases of misbehaviour or abnormal context.
- **Amobility** is the operator of the autonomous vehicles in the Copenhagen site and is handling all daily operation of the vehicles and everyday contact with the end users. Amobility is leading the T4.2 and will be coordinating and developing the testing framework for the service validation and prototyping of the services in the vehicles. As an operator Amobility provides the autonomous shuttles that will be used for the AVENUE pilot activities and its facilities for performing data capture activities, deploying the service and testing its performance through short or longer evaluation periods. The shuttles that will be used to test and prototype the service are located in Copenhagen,Nordhavn and Oslo, Omrøya. Once the prototyping and testing framework has been developed and tested, the prototyping session will be coordinated on the other AVENUE sites, if possible. Hence via TPG, SLA and KEOLIS.

3.3 Technical requirements

The “Enhance the sense of security and trust” service exhibits a variety of advantages, which include the onsite intelligence of the platform using low cost devices that can decide with minimal latency whether the cloud side of the system must be notified, saving thus network/cloud resources. Moreover, the onboard side analytics can confirm or reject a video clip as true or false positive respectively, provides the relevant information to the end user, allowing him to focus his/her attention on crime incidents. In addition, the identification components will ease the task of authorities contributing to the timely apprehension of any offender.

In the next sections, the indicative technical specifications of the system are presented to host the algorithms that are developed based on RGB, audio and depth sensors for the embedded and cloud side of the surveillance platform. In this context, various analysis approaches for early petty crime detection are adopted, such as

1. Video analytics
2. Depth analytics
3. Audio analytics

The embedded system is

- Easy to install and use (normally less than an hour). Adaptable to the network conditions.
- Supports a multitude of video/audio formats and codecs (98 codecs; including all the most widely used ones).
- Supports standalone operation when connection with the cloud is not an option.
- Allows real time re-configuration of the system's components.

Certain sensors need to be installed in the autonomous vehicles which include:

- Embedded camera support
- USB cameras support
- IP cameras support
- USB, IP microphones support

The sensors of the embedded system can include various types, have enhanced interoperability and support integration with legacy cameras. When an event is detected a clip is generated. The streams will be cut to 1 minute packages and will be analysed.

The system utilizes standard of-the-shelf **cameras** and **microphones**, so no specialized protocols and/or ports are required. In most cases the sensors will be directly connected with the Embedded System through a USB port, whereas IP cameras could also be supported, with the latter being connected through Ethernet or Wi-Fi with the Embedded System. The following tables present the detailed characteristics of indicative sensors that can be used. At this point it has to be noted that since the system can support a number of different cameras and microphones, the minimum requirements are presented and not a specific model.

Sensor/Device: Video Sensing

Name	The system supports most of-the-shelf cameras
Short Description	The sensor can be utilized to detect abnormal events in the indoor environment of the autonomous shuttles.
Measurement	The sensor must be capable of acquiring colour images and should have a night mode. The Sensor will be able to acquire data that will be fed to the Video analytics algorithms for further analysis. The sensor is attached to a PC using the USB-interface. IP cameras are also supported.
Functionality	The sensor will be part of the system in order to detect abnormal events (e.g., theft, fighting, anti-social behaviour) on the autonomous shuttles. Moreover, in case an abnormal event is detected, data will be sent to the AVENUE platform in order to identify the suspect(s).

Indicative specifications are presented next. Detailed specifications will be derived based on discussions with Navya, the manufacturer of the autonomous vehicles. The dimensions and the weight of the sensor vary and the mounting can be made manually. In terms of resolution, the sensor should provide minimum HD resolution images (1280 × 720 pixels) at 24 frames per second and the operating temperature range applies (e.g., 5 to 35 degrees Celsius). The sensor will be powered via the USB 2.0 or Power over Ethernet (PoE) connection. An IP Camera should support widely used streaming protocols (e.g RTSP).

The RGB video stream should be 24-bit HD (ideally FullHD 1080p) resolution. The data rate is at minimum 30 Hz but it highly depends on the actual application use for real-time performance. The data stream is continuous and can be acquired also as required (or on demand). The API and the SDK of the camera is flexible enough to acquire the images as soon as they are required on the aforementioned maximum data epoch rate.

The transmission frequency should be 30 Hz. No specific software is required. The sensor acquires data that could be used to identify a person. However, within the system special countermeasures will be taken into account in order to address all ethical, legal and privacy issues.

Sensor/Device: Audio Sensing

Name	The system supports most of-the-shelf microphones
Short Description	The sensor can be utilized on the interior of public transportation (i.e. autonomous shuttles)
Measurement	The sensor must be capable of acquiring audio at 16-bit audio at a sampling rate of 16 kHz. The sensor is attached to a PC using the USB-interface.
Functionality	The sensor will be part of the system in order to detect abnormal events (e.g., screaming, breaking glass) on the indoor environments of autonomous shuttles. Moreover, in case an abnormal event is detected, data will be sent to the AVENUE platform in order to notify the related stakeholders.

Indicative specifications are presented next. Detailed specifications will be derived based on discussions with Navya, the manufacturer of the autonomous vehicles. The dimensions and weight of the sensor vary and the mounting should be made manually. The operating temperature range applies (e.g., 5 to 35 degrees Celsius). The data connection is USB 2.0 through which the powering is performed. The data format is WAVE format. The sensor acquires data that could reveal a person's identity. However, within the system special countermeasures will be taken into account in order to address all ethical, legal and privacy issues.

Other requirements

A PC to host the system and all developed algorithms is required. Power supply and data connectivity are also required in this case.

3.3.1 Technology

In the following sections, the research, involving algorithms and experiments conducted, is presented for the "Enhance the sense of security and trust" service. As depicted in Figure 2, the first layer of sensors connects to the Hardware Abstraction Layer (HAL). The HAL implements the IP and the USB protocol supporting IP and USB camera, respectively. The input data is converted and transformed in a compatible format and passed into the analytics algorithms. The prediction and a short video clip (if required) are then transferred via the API endpoints into the cloud. The operator has access to the data and acts accordingly.

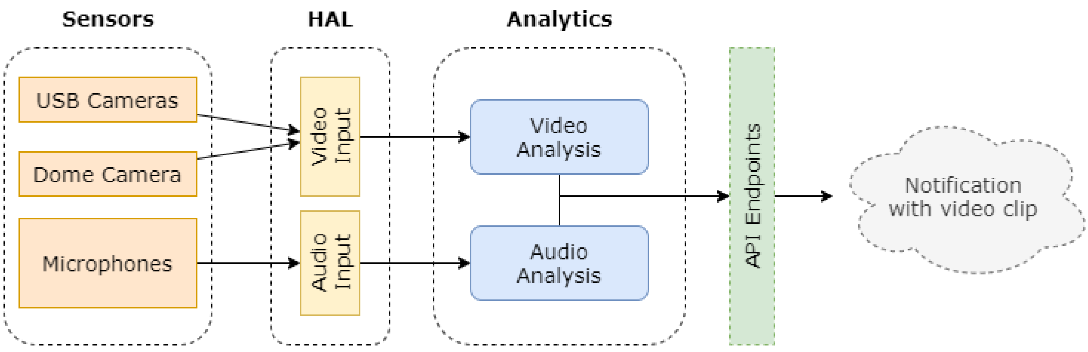


Figure 2. High level overview of the "Enhance the sense of security and trust" service

In this context, various analysis approaches for early petty crime detection are implemented, such as:

1. Video analytics (described at Section 2.4.3.1.1)
2. Audio analytics (described at Section 2.4.3.1.2)

Each technique is presented and explained, along with its results and its major shortcomings.

3.3.1.1 Video analysis

This section focuses on the services that are going to be deployed in the autonomous shuttle, using images from video feeds. Specifically, we describe the datasets that have been used and the algorithms for the image-based abnormal event detection starting from the pre-processing steps to the classification deep neural network modules.

For video analysis, three different approaches were implemented: (a) a Pose Long Short-Term Memory (LSTM) Classifier, (b) a Spatiotemporal Autoencoder and (c) a Spatiotemporal LSTM Classifier.

Background

Automatic awareness of human actions and its interaction with the environment has become a prominent study area in recent years. To perform such a demanding mission, many scientific areas rely on modeling human activity in its various dimensions (emotions, relational attitudes, actions, etc.). In this context, the identification of a person's activity tends to be necessary to the comprehension of specific acts. Thus, a great interest has been granted to human action recognition, especially in real-world environments. There are various attempts on human action recognition based on RGB video and 2D/3D skeleton data. The RGB video-based action recognition methods[1] [2] [3] [4] mainly focus on modeling spatial and temporal representations from RGB frames and temporal optical flow.

Despite RGB video-based methods have achieved promising results, there still exist some limitations, e.g., background clutter, illumination changes, appearance variation, and so on. 3D skeleton data represents the body structure with a set of 3D coordinate positions of key joints. Since skeleton sequence does not contain color information, it is not affected by the limitations of RGB video. Such robust representation allows to model more discriminative temporal characteristics about human actions.

Moreover, Johansson et al.[5] have given an empirical and theoretical basis that key joints can provide highly effective information about human motion. Besides, the Microsoft Kinect[6] and advanced human pose estimation algorithms[7] make it easier to gain skeleton data. For skeleton based action recognition, the existing methods explore different models to learn spatial and temporal features. Song et al.[8] employ a spatial-temporal attention model based on LSTM to select discriminative spatial and temporal features. The Convolutional Neural Networks (CNNs) are used to learn spatial-temporal features from skeletons in[9] [10] [11]. Yan et al. [12] employ graph convolutional networks (GCN) for action recognition. They propose to utilize the graph neural network and LSTM to represent spatial and temporal information, respectively. In short, all these methods are trying to design an effective model that can identify spatial and temporal features of skeleton sequence. Nevertheless, how to effectively extract discriminative spatial and temporal features is still a challenging problem.
In this section, we propose three different approaches for the detection of anomalous events: (a) a Pose Long Short-Term Memory (LSTM) Classifier, (b) a Spatiotemporal Autoencoder and (c) a Spatiotemporal LSTM Classifier.

---

[1] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014

[2] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In ECCV, 2016

[3] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In ICCV, 2015

[4] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. Rgb-d-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding, 2018

[5] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. Perception & Psychophysics, 1973

[6] Zhengyou Zhang. Microsoft kinect sensor and its effect. IEEE Multimedia, 2012.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In CVPR, 2017

[8] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In AAAI, 2017

[9] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In ACPR, 2015

[10] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Cooccurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In IJCAI, 2018

[11] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In CVPR, 2017

[12] Sijie Yan, Yuanjun Xiong, Dahua Lin, and xiaoou Tang. Spatial temporal graph convolutional networks for skeletonbased action recognition. In AAAI, 2018.

Datasets

The experiments require a large amount of data due to the nature of the machine learning algorithms that CERTH developed. The training process of the models depends on the amount of input data, in order to be accurate and satisfy the service requirements. Most of the data were not available, thus several data capture sessions were performed by simulating the shuttle environment as close as possible. A data capture in CERTH facilities was conducted as an initial approach, whereas a video from the indoor environment of Navya shuttle was examined. Moreover, a data capture in TPG facilities from indoor environment of the shuttle as a real-environment approach was also conducted.

At the first data capture at CERTH facilities, 13 different scenarios were simulated using two different camera perspectives for each one. The dataset contains 6650 frames that, in conjunction with some augmentation techniques described later, were sufficient to train the LSTM Classification via Pose Estimation experiment and obtain decent results. During the second data capture performed in TPG depot, 29 video sequences were captured with 46127 frames, demonstrating real conditions in an autonomous vehicle. The merging of these two datasets led to substantially better results and allowed to perform more experiments. Samples from NTU-RGB[13] were dataset further included, which helped to access the performance on unknown data and fine tune the model to generalize. Also, samples from P-REACT[14] dataset were used to verify the results and implement the spatiotemporal autoencoder. Most of the collected data were annotated by a cross-platform utility specifically developed for this purpose. This tool was designed to improve productivity as it automates many procedures related to the time-consuming labelling process. The utility integrates the service's critical components, such as the pose estimation backend and uses the same data format and structure. The simple and intuitive graphical user interface supports batch navigation, per frame or skeleton labelling and class switching and was designed to improve productivity as it automates many procedures. All five of the datasets are depicted in Figure 3, while the cross-platform utility specifically developed for this purpose is depicted in Figure 4.

[13] Shahroudy, Amir, et al. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[14] Dimitriou, Nikolaos, et al. "An Integrated Framework for the Timely Detection of Petty Crimes." 2017 European Intelligence and Security Informatics Conference (EISIC). IEEE, 2017.


CERTH LAB
TPG DEPOT
P-REACT
NTU-RGB-D
HOLO + TPG SHUTTLE

Figure 3. Dataset samples with abnormal events that showcase the use cases (fighting, aggression, bag-snatching, and vandalism). The red section contains simulated data in lab, the green section depicts captured data from TPG shuttles, the blue section shows scenarios from the P-REACT dataset, the grey section indicates additional data imported from the NTU-RGB dataset and the orange section additional data captured from HOLO and TPG shuttles.

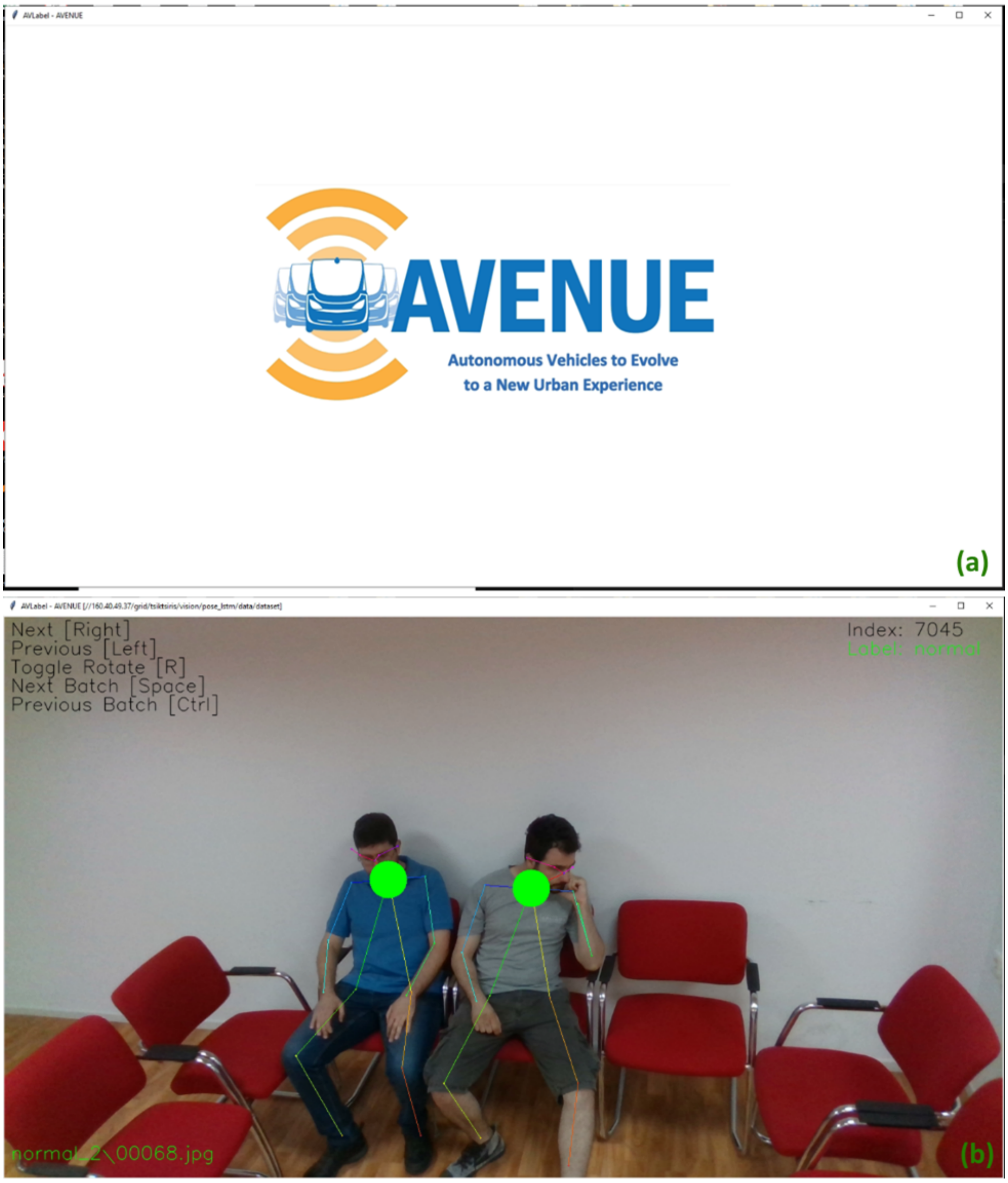


Figure 4. Screenshots of the utility: (a) Initialization screen (b) Pose labelling using mouse clicks on the skeleton's circle and keyboard navigation

LSTM Classification via Pose Estimation

The pipeline of this approach consists of 4 stages as shown in Figure 5. In stage one, the pose of each person in the frame is extracted (15 keypoints). In the second stage, a skeleton tracking algorithm is performed, associating persons across multiple frames. Regarding the third stage, the detected and tracked human body key-points are represented as trajectories and during the fourth stage they are "fed" into a Multi-Layer Classifier which classifies each action into normal or abnormal.

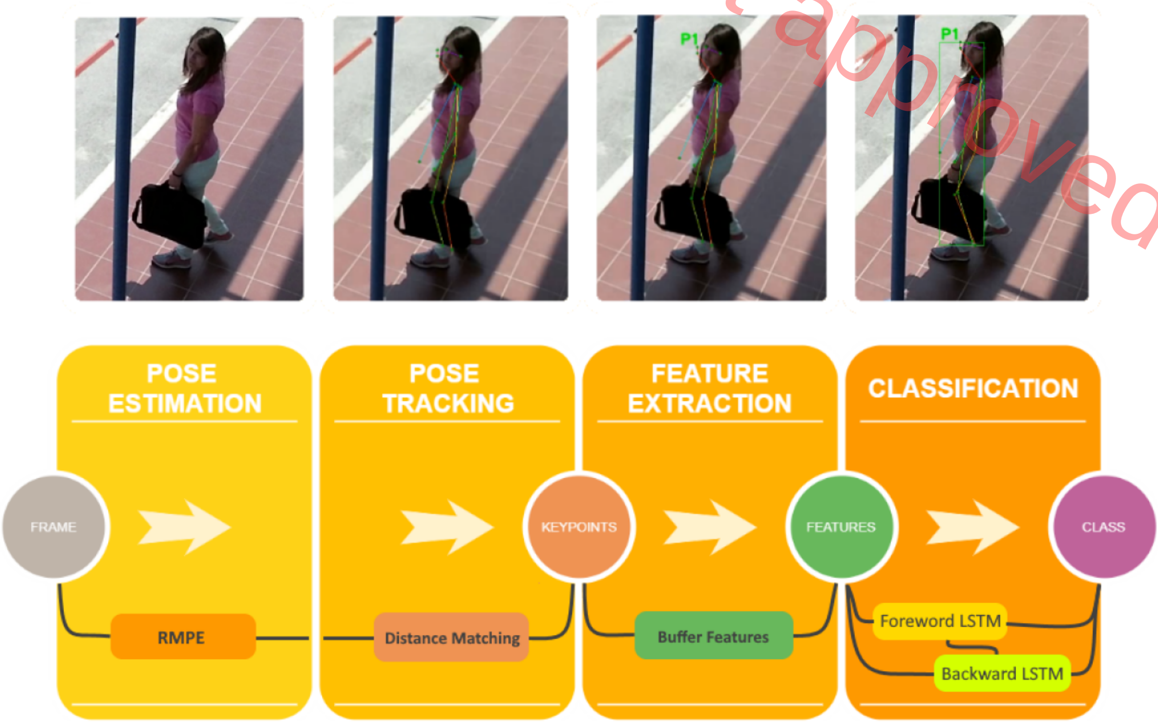


Figure 5. Pipeline of the Pose Estimation Classification

Pose Estimation: For the first stage, we are using a custom implementation of the Regional multi-person pose estimation by Fang et al.[15] for training (better accuracy) and OpenPose[16] for testing (higher performance). The integrated implementation uses VGG19, a convolutional neural network model proposed by K. Simonyan and A. Zisserman[17]. We are using this model to improve the accuracy of the skeleton extraction for the training process and further perform data augmentation without sacrificing the data integrity. We generate noisy data with variable intensities, based on the extracted data from the backend, and we combine these data with the original ones as an augmentation technique. Extensive tests indicated that the model generalizes better and the accuracy improves. Although we initially implemented AlphaPose using VGG for training our model, multiple tests shown that we could use it for evaluation too, when specific parameters (resolution, heatmaps etc.) are tuned. We apply the pose estimation framework in the presence of inaccurate human bounding boxes. The generated pose proposals are refined by parametric pose non-maximum suppression to obtain the estimated human poses. In this stage, 17 different human body keypoints are detected and the number of people in each frame is obtained. The number of N frames for the feature generation along with the evaluation accuracy are depicted in the graph below (Figure 5). As we can see, a buffer size (window size) of 5 frames achieved the best accuracy on the evaluation test. Higher values may result in lower accuracy as the tracker may fail to consistently detect people when the shuttle is overcrowded. We are currently investigating some optimizations of the tracker, in order to further increase the size of the buffer.

[15] Fang, Hao-Shu, et al. "Rmpe: Regional multi-person pose estimation." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[16] Cao, Zhe, et al. "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields." arXiv preprint arXiv:1812.08008 (2018).

[17] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

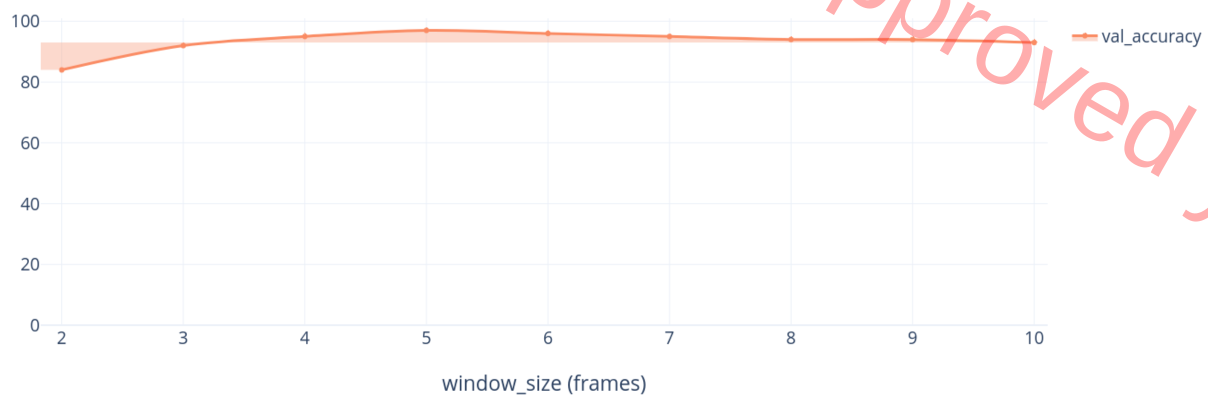


Figure 6. Performance evaluation across different buffer sizes

Tracking: An important step is to match cross-frame poses and form pose flows (tracking). Also, a novel pose flow non-maximum suppression is applied to reduce redundant pose flows and re-link temporal disjoint ones. This is an important step that associates poses indicating the same person, across multiple frames.

We implemented an online skeleton tracking algorithm based on distance and other heuristics, in order to meet the performance requirements of the real-time service. The algorithm is sorting the skeletons based on the distance between neck and image center, from small to large.

A skeleton near the center will be processed first and be given a smaller human id. Later on, each skeleton's features will be matched between the current and the previous frame.

Figure 7. Skeleton matching across two different frames (blended)

The distance matrix (or cost) between the skeleton joints is the main criterion for the matching function. Skeletons with the small distance are matched between the frames and are given the same id (Figure 7). In some cases, the skeleton detection framework might fail to detect a complete human skeleton from the image due to the restricted field of view of the camera in the autonomous vehicle, as depicted in Figure 8.

Figure 8. Examples indicating the restricted field of view for the camera sensor

This may cause some blanks in the joint positions, which should be filled with some values in order to maintain a fixed-size feature vector for the following feature classification procedure.

We evaluated some solutions for this issue:

- **Solution 1:** Discard this frame. However, the algorithm would never be able to detect the action when the person is standing sideways and not facing the camera.
- **Solution 2:** Fill in the positions with some value outside a reasonable range. Theoretically, when the classifier is strong enough, this method could work.
- **Solution 3:** Fill in a joint's position based on its relative position in the previous frame with respect to the neck.

In order to solve this issue, solution 3 was implemented, but the classifier’s performance was degraded in some test cases. After extensive tests, we noticed that a previous joint position might be missing too, being replaced by the estimation of our algorithm. This led to “stuck” joints across various frames and confused both the tracker and the classifier.

To overcome this issue, we are using a default “idle” pose as an example for our algorithm. When a previous joint is missing, the value being replaced is relative to the default example pose (Figure 9). We chose a person sitting as the default, because it is the most common for the passengers in the AV.

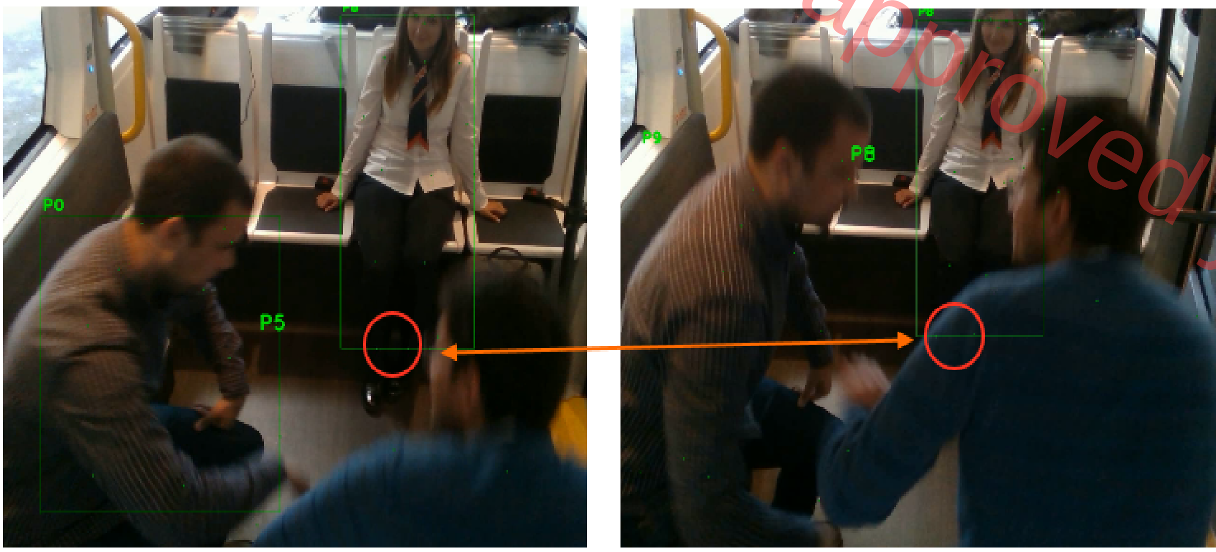


Figure 9. The two leg joints are being reconstructed (right) by their relative location in the previous frame (left)

Feature Extraction: Regarding the third stage, the detected and tracked human body key-points are converted into features and forwarded to an LSTM Neural Network. For extracting features, we store every person's skeleton data into a circular (ring) buffer deque (double-ended queue) of N frames (window size) into the Feature Generator class.

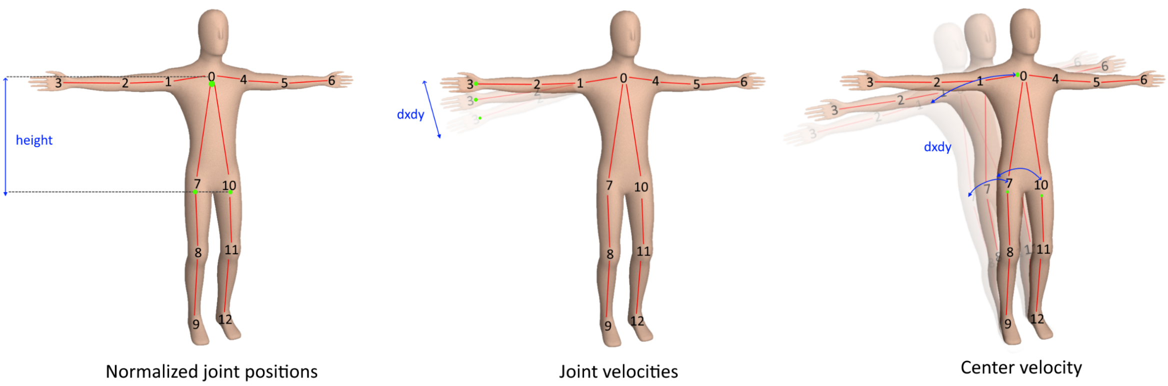


Figure 10. Representation of the extracted features

We consider the buffer as invalid if the newest appended skeleton does not contain at least the neck (Point 0) or one of the thigh bones (Point 7 or 10) shown in Figure 10, as the height of the skeleton (used for normalizing features) cannot be calculated. The feature extraction process occurs when the buffer is full.

Table 3. Features extracted from pose classification

Pose Classification – Feature Extraction	
Xs	A direct concatenation of joints positions of the N frames.
H	Average height of the skeleton of the previous N frames. This height equals the length from Neck to Thigh. Used for normalizing features.
X	Normalized joint positions [Xs – mean(Xs)]/H
Vj	Velocities of the joints {X[t]-X[t-1]}

Vc	Velocity of the center {sum(Xc[t]-Xc[t-1])} (**10x** weight)

The number of N frames for the feature generation along with the evaluation accuracy are depicted in Figure 9. A buffer size (window_size) of 5 frames achieved the best accuracy on the evaluation test. Higher values may result in lower accuracy as the tracker occasionally fails to consistently track people when the shuttle is overcrowded.

Classification: The LSTM model is capable of binary or multi-class softmax classification. It contains three hidden layers of size (32 × 64) with the rectified linear unit (ReLU) activation function. An overview is shown in Figure 11.

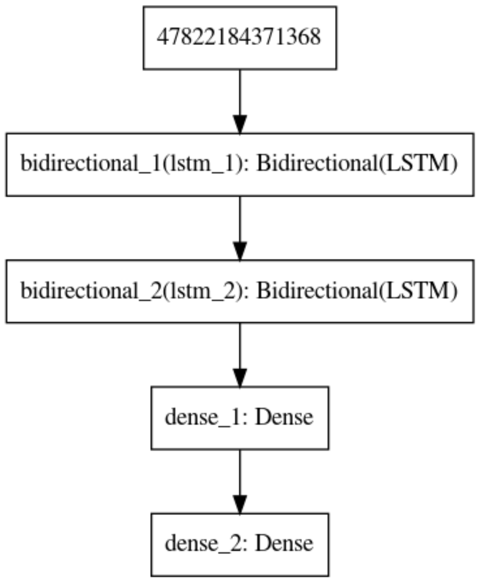


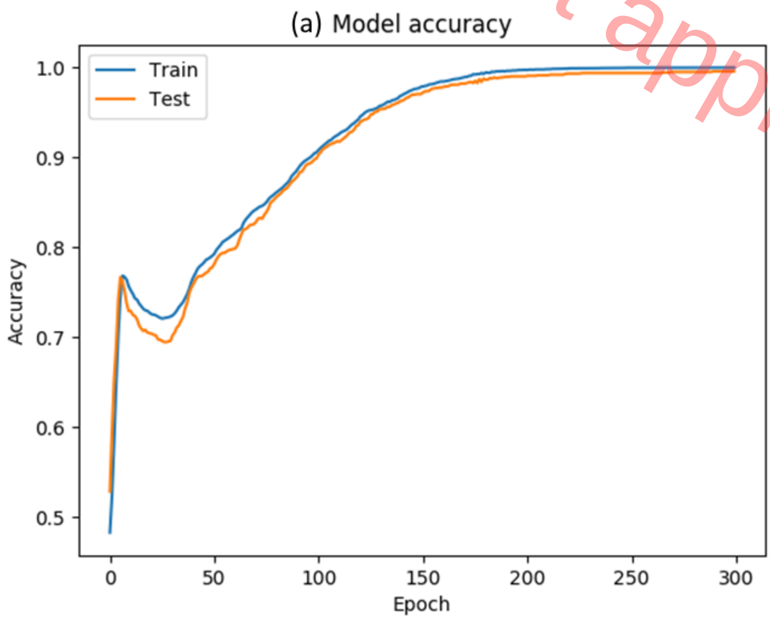
Figure 11. Model Overview

The model is trained end-to-end and regularised so that it distils the most compact profile of the normal patterns of training data and effectively detects abnormal events (Figure 12).

Figure 12. Abnormal event detection (passengers are fighting)

We applied a Principal Component Analysis (PCA) procedure to reduce the 314 initial features to 50 principal components. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors (each being a linear combination of the variables and containing n observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables. After the PCA procedure, we achieved a sum of 0.9981\% eigenvalues. The model converges faster and the accuracy improves. Due to the highly imbalanced classes, we compiled our model with a custom **weighted** categorical cross-entropy function for loss calculation, and a weighted categorical accuracy method in order to acquire accurate metrics.

The evaluation performance of our model is depicted below (Figure 13). After 50 epochs, our model achieved a 96.22% accuracy. The time cost for feature extraction and classification is less than 0.05s per frame for the classifier, since the model is relatively shallow.



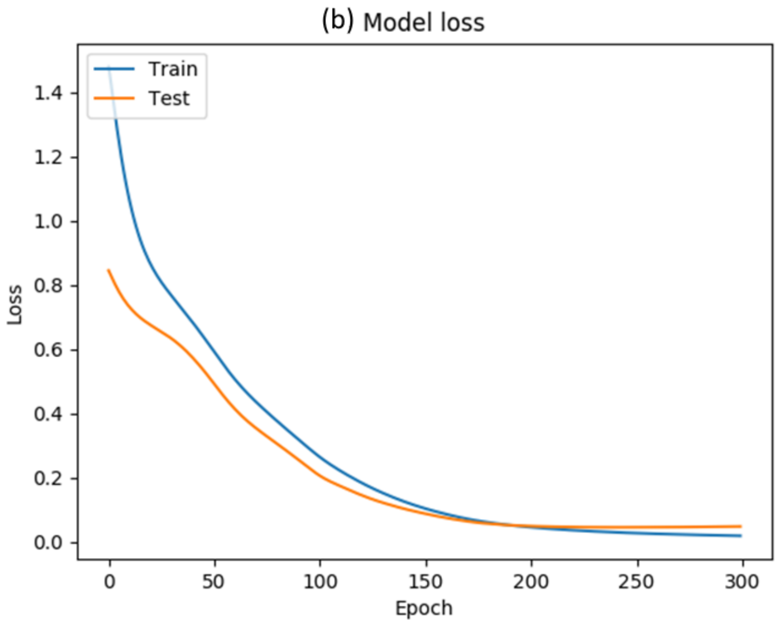


Figure 13. Model performance evaluation – (a) Training accuracy and (b) Training loss metrics

Spatiotemporal Autoencoder

The second approach is based on the principle that the most recent frames of video will be significantly different than the older frames, in case of an abnormal event. Our goal, inspired by Lu et al.[18] , is to train an end-to-end model consisting of both a spatial feature extractor and a temporal encoder-decoder that combined learn the temporal patterns of the input volume of frames. So as to minimize the reconstruction error between the input and the output video volume reconstructed by the learned model, we trained our model using video volumes of only normal scenes. After our model's proper training, we expect to have low reconstruction error in a normal video volume as opposed to a video volume containing abnormal scenes. Finally, our system will be able to detect the occurrence of an abnormal event, by thresholding on the error produced by each testing input volume.

Architecture: There are two stages that form an autoencoder: encoding and decoding. Autoencoders set the number of encoder input units to be less than the input; thus, they were first used to reduce dimensionality. Usually, unsupervised back-propagation is used for training, helping the reconstruction error of the decoding results from the original inputs to decrease. Generally, an autoencoder can extract more useful features when the activation function is non-linear rather than some common linear transformation methods, such as Principal Component Analysis (PCA).

Spatial Convolution: In a deep CNN, the main objective of convolution is to extract information from the input frame. The convolution process maintains the spatial relations of pixels by using kernels to extract low level features. In raw mathematics, the convolution operation performs dot products across filters of partial input. Supposing an nxn input layer, followed by a convolutional layer, then if we use an $m \times m$ filter W, the output size will be $(n - m + 1) \times (n - m + 1)$. Through the training stage, a CNN learns the values of these filters by itself, although some parameters such as the filter size and the number of layers still need to be defined. The larger the number of filters used, the more information that gets extracted and the better the network generalizes. Yet, there is a trade-off and balance is a critical factor when it comes to the number of filters used, as more filters would impact the performance negatively and require more resources.

Preprocessing: At this stage, our task is to convert raw data into aligned and acceptable input for the model. To do so, each frame extracted from the raw videos is then resized to 64 ×64. To ensure that the input images are all on the same scale, the pixel values are scaled between 0 and 1 and each frame subtracted by its global mean image for normalization. The mean image is calculated by averaging the pixel values at each location of each frame in the training dataset. After that, the images are converted to grayscale to reduce dimensionality. Finally, the processed images are then normalized to have zero mean and unit variance. As mentioned before, we use video volumes as input to our model, where each volume consists of 10 consecutive frames with various skipping strides (Figure 114. As the number of parameters in this model is large, we also need a large amount of training data. Following Lu's practice, to increase the size of the training dataset, we perform data augmentation in the temporal dimension. To generate these volumes, we concatenate frames using sequences, namely being stride-1, stride-2, and stride-3. For example, the stride-1 sequence consists of frame numbers {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, whereas the first stride-2 sequence contains frame numbers {1, 3, 5, 7, 9, 11, 13, 15, 17, 19}. Now the input is ready for model training.

[18] Lu, Yiwei, et al. "Future Frame Prediction Using Convolutional VRNN for Anomaly Detection." 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019.

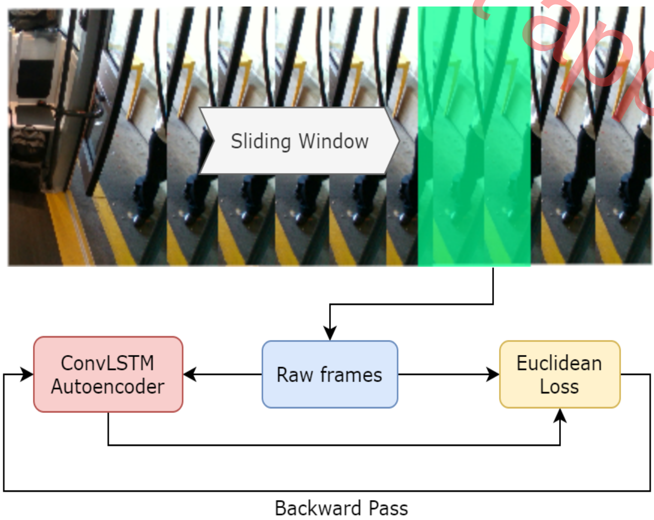


Figure 14. Preprocessing and training pipeline

Feature Learning: In order to learn the regular patterns in training videos, we propose a convolutional spatiotemporal autoencoder. Our architecture consists of two parts — a spatial autoencoder for learning spatial structures of each video frame, and a temporal encoder-decoder for learning temporal patterns of the encoded spatial structures. As illustrated in Figure 14, the spatial encoder and decoder have two convolutional and deconvolutional layers respectively, while the temporal encoder is a three-layer convolutional long short-term memory (LSTM) model. Convolutional layers are well-known for their superb performance in object recognition, while the LSTM model is widely used for sequence learning and time-series modelling and has proved its performance in applications such as speech translation and handwriting recognition.

Autoencoder: There are two stages that form an autoencoder: encoding and decoding. Autoencoders set the number of encoder input units less than the input; thus, they were first used to reduce dimensionality. Usually, unsupervised back-propagation is used for training, minimizing the reconstruction error of the decoding results from the original inputs. Generally, an autoencoder can extract more useful features when the activation function is non-linear rather than some common linear transformation methods, such as PCA.

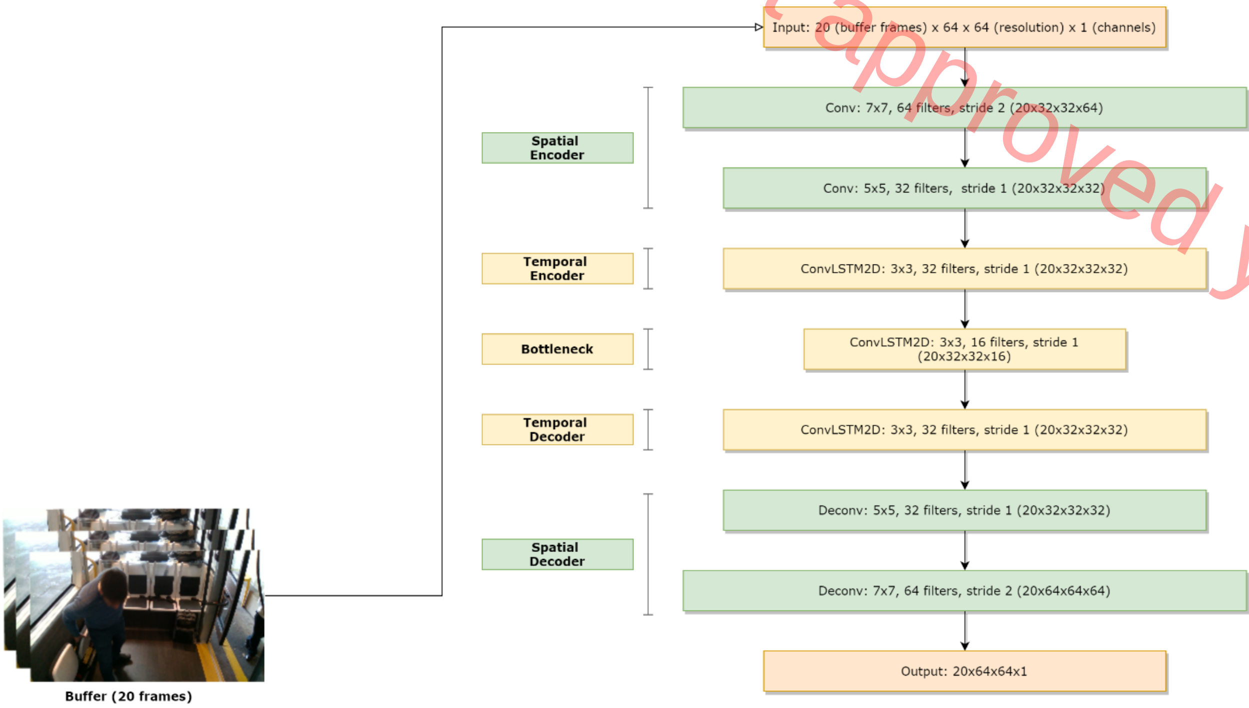


Figure 15. Autoencoder Model Architecture

Spatial Convolution: The primary purpose of convolution in a convolutional network is to extract features from the input image. Convolution can preserve the spatial relationships between pixels by using small squares of input data to learn image features. Mathematically, convolution performs dot products between the filters and local regions of the input. Assuming that we have some n x n square input layer, followed by the convolutional layer, then if we use a m x m filter W, the convolutional layer output will be of size $(n - m + 1) \times (n - m + 1)$.
During the training process, a convolutional network learns the values of these filters on its own, although parameters such as the number of filters, filter size, the number of layers before training still need to be specified. The larger the number of filters used, the more image features get extracted and the better the network becomes at recognizing patterns in unseen images. However, balance is key when it comes to the number of filters used, as more filters would add to computational time and exhaust memory faster.

Recurrent Neural Network (RNN): In a traditional feedforward neural network, we assume that all inputs (and outputs) are independent of each other. However, in tasks involving sequences, learning temporal dependencies between inputs are important, as e.g., a model of word predictor should be able to derive information from the past inputs.

An RNN works just like a feedforward network, except that the values of its output vector are influenced not only by the input vector but also on the entire history of inputs. In theory, RNNs can make use of information in arbitrarily long sequences, but in practice, due to vanishing gradients, they are limited to looking back only a few steps.

Long Short-Term Memory (LSTM): To overcome this problem, a variant of RNN is introduced: a Long Short-term Memory (LSTM) model (Figure 16) that incorporates a recurrent gate called forget gate. With the new structure, LSTMs prevent backpropagated errors from vanishing or exploding. Therefore, LSTMs can work on long sequences and can be stacked together to capture higher level information.

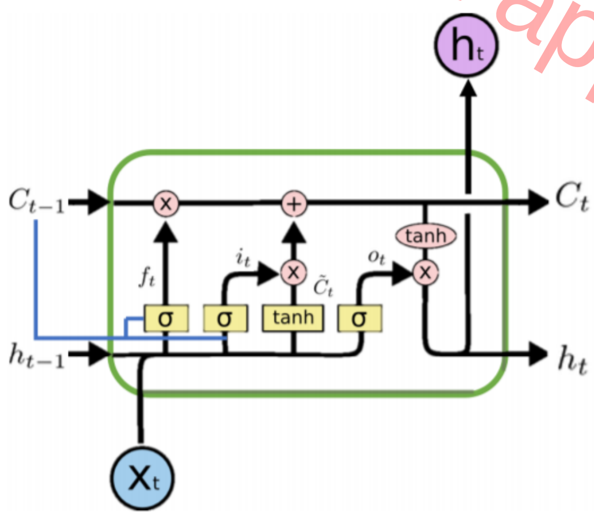


Figure 16. The structure of a typical LSTM unit. The blue line represents an optional peephole structure, which allows the internal state to look back (peep) at the previous cell state Ct−1 for a better decision

Convolutional Long Short-term Memory (ConvLSTM): The Convolutional Long Short-term Memory (ConvLSTM) model, considered a variant of the LSTM architecture, was introduced by Shi et al. in[19] and has been recently utilized by Patraucean et al.[20] for video frame prediction. Compared to the usual fully connected LSTM (FC-LSTM), ConvLSTM, as shown in Figure 17, has its matrix operations replaced with convolutions. ConvLSTM requires fewer weights and yields better spatial feature maps, by using convolution for both input-to-hidden and hidden-to-hidden connections.

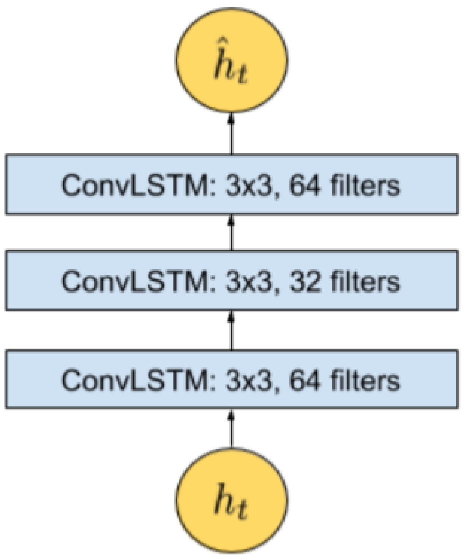


Figure 17. The zoomed-in architecture at time t, where t is the input vector at this time step. The temporal encoder-decoder model has 3 convolutional LSTM (ConvLSTM) layers.

[19] Xingjian, S. H. I., et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." Advances in neural information processing systems. 2015.

[20] Patraucean, Viorica, Ankur Handa, and Roberto Cipolla. "Spatio-temporal video autoencoder with differentiable memory." arXiv preprint arXiv:1511.06309 (2015).

Regularity Score: We define the reconstruction error as the Euclidean distance across an input and a reconstructed frame. Specifically, the following equations describe the reconstruction error, where t denotes the frame in a sequence:

$$e(t) = \left| x(t) - f_w(x(t)) \right|^2 \qquad (1)$$

where f_w is the learned weights by the spatiotemporal model. The irregularity score $s_a(t)$ is calculated by scaling between 0 and 1. Eventually, the regularity score $s_r(t)$ can be derived by subtracting the reconstruction score from 1:

$$s_a(t) = \frac{e(t) - e(t)_{min}}{e(t)_{max}} \qquad (2)$$

$$s_r(t) = 1 - s_a(t) \qquad (3)$$

Once the model is trained, its performance can be evaluated by feeding in testing data and checking whether it can detect abnormal events while maintaining a low false alarm rate. For a better comparison, we used the same formula as Hasan et al.[21] to calculate the regularity score for all frames. Our only difference is that the learned model is of a different kind. The reconstruction error of all pixel values in a frame of the video sequence is taken as the Euclidean distance between the input frame and the reconstructed frame (Figure 18). The reconstruction error or cost of a frame sequence is calculated as the difference between the ground truth (original frames) and the reconstructed frames (prediction – model output). The frame sequence is flagged as “abnormal” if the reconstruction cost exceeds a certain threshold. An example of a prediction with a low regularity score is shown in Figure 19.

Thresholding: Using a threshold on the reconstruction error, we are able to determine whether a video frame is normal or abnormal. A predefined threshold is not a robust method since our solution should operate in real-time and support multiple camera sensors as we mentioned earlier in the design principles. A fixed threshold value can alter the sensitivity of the event detection, rendering it inappropriate in some scenarios. In addition, a wrong threshold can prevent the detection of certain abnormal events or produce false positives. In order to solve this issue, we introduce a variable thresholding technique in order to find the optimal value in real-time. The initialization procedure now includes a “warm-up” session, in which we aggregate the individual regularity score of each frame. During that session, no detections are performed, as we consider the events as regular. As the buffer continues to fill, we are able to calculate the average reconstruction error and provide a threshold value tailored to the specific conditions. Figure 19 indicates an abnormal scenario with the aforementioned metrics.

[21] Hasan, Mahmudul, et al. "Learning temporal regularity in video sequences." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

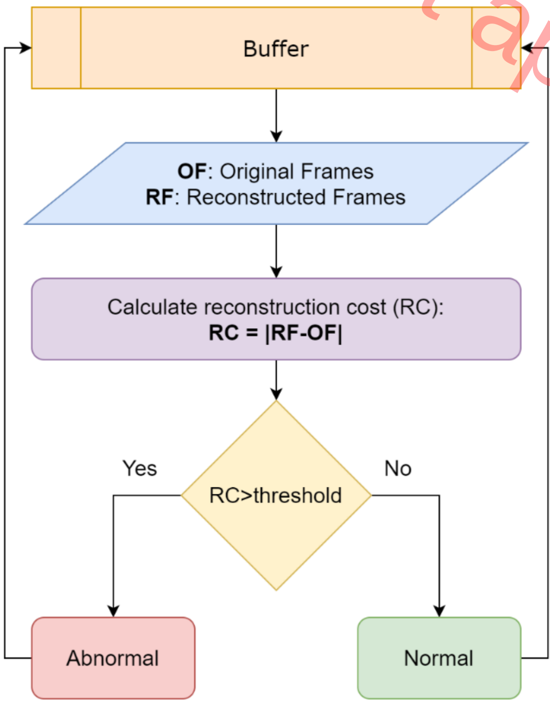


Figure 18. Architecture overview of the autoencoder approach

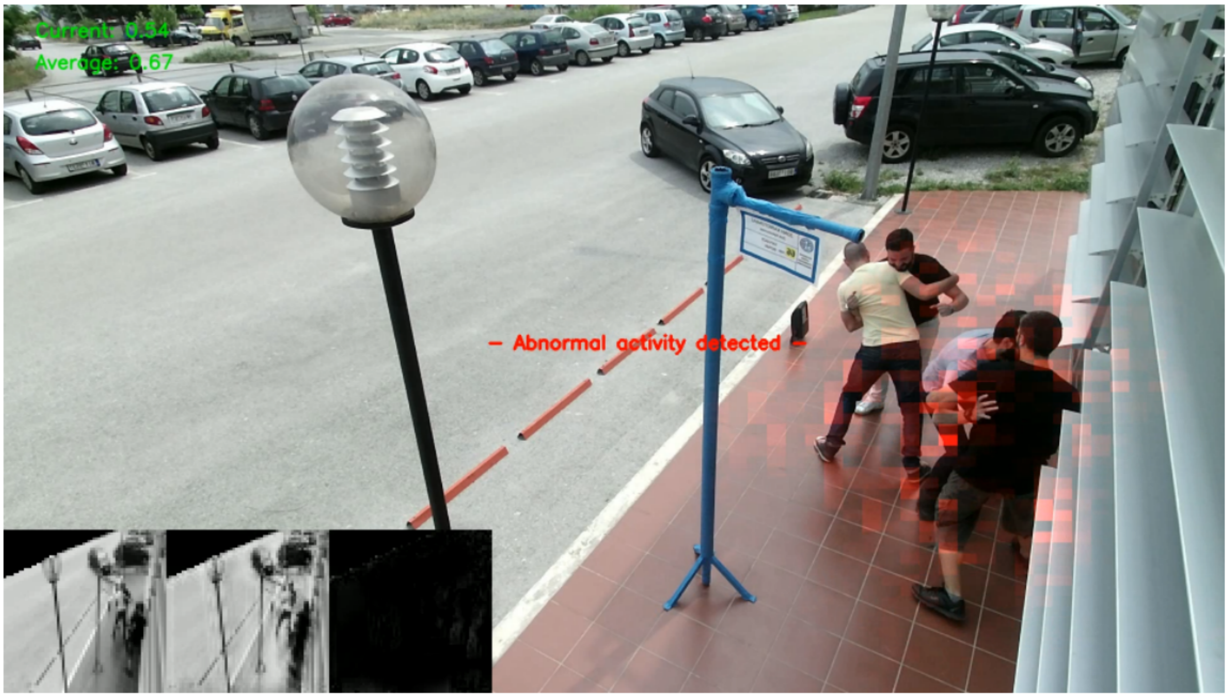


Figure 19. Outdoor group fighting scenario on a simulated bus stop. Green metrics at top-left indicate the current and the average regularity score. A lower regularity score indicates that the predicted reconstruction is not accurate, since our model did not learn such an event. Note that the current score is much lower than the average, triggering an abnormal notification.

Hybrid LSTM Classification

Even if our dataset consists of thousands of data, we will get some anomalies on certain occasions. So now humanly it is possible to manually go through the anomaly outputs and flag some of them as false positives. Therefore, we can let the previous autoencoder neural network model act as a **High Recaller**.

Semi-supervised Learning: The threshold is decreased so that almost all the actual anomalies are getting detected (high recall) along with other false positive anomalies (low precision). To achieve the semi-supervised approach, we designed a new model which includes the previous Encoder and an LSTM which acts as a classifier.

Figure 20. Example of a prediction with a lower regularity threshold

In real-time inference the anomalies predicted by the high recaller model (autoencoder neural network) are sent through the false positive reduction model (hybrid model). This combination of neural networks (Figure 19) should provide a deep neural network model with high recall and high precision.

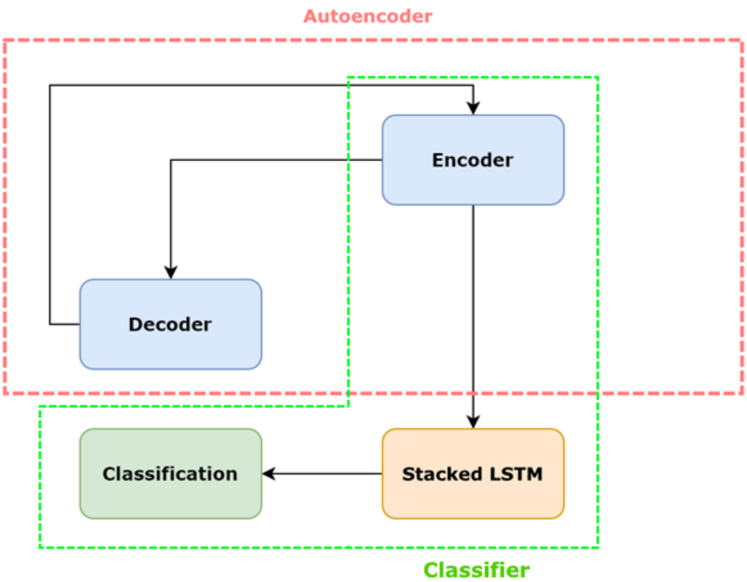


Figure 21. Model architecture of the hybrid model. The red container contains components of the previous autoencoder approach. The green components indicate the new hybrid model which acts as a classifier

Training: The training process of the new experiment consists of 3 stages and it is depicted in Figure 22. At first, the autoencoder model (encoder + decoder) is being trained with unsupervised data to learn regularity. In the second stage the encoder weights are transferred to the hybrid model. The encoder's layers are marked as non-trainable. Finally, we perform supervised training using only the LSTM classifier.

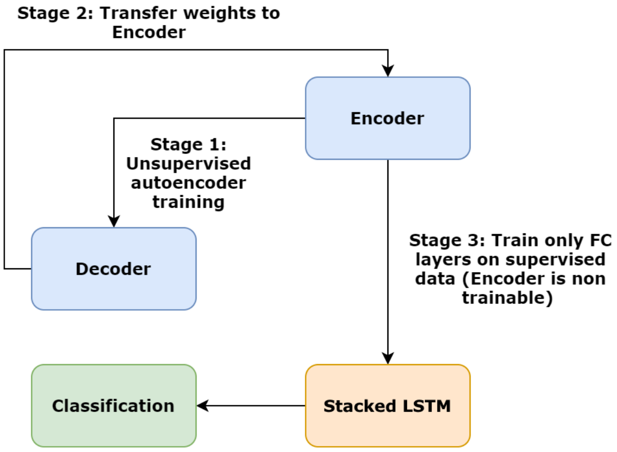


Figure 22. Pipeline of the hybrid training procedure

We trained the hybrid model for 20 epochs, using a 64/16/20 train/validation/test ratio. The loss and the accuracy are depicted in Figure 23 and the final model accuracy is 98%.

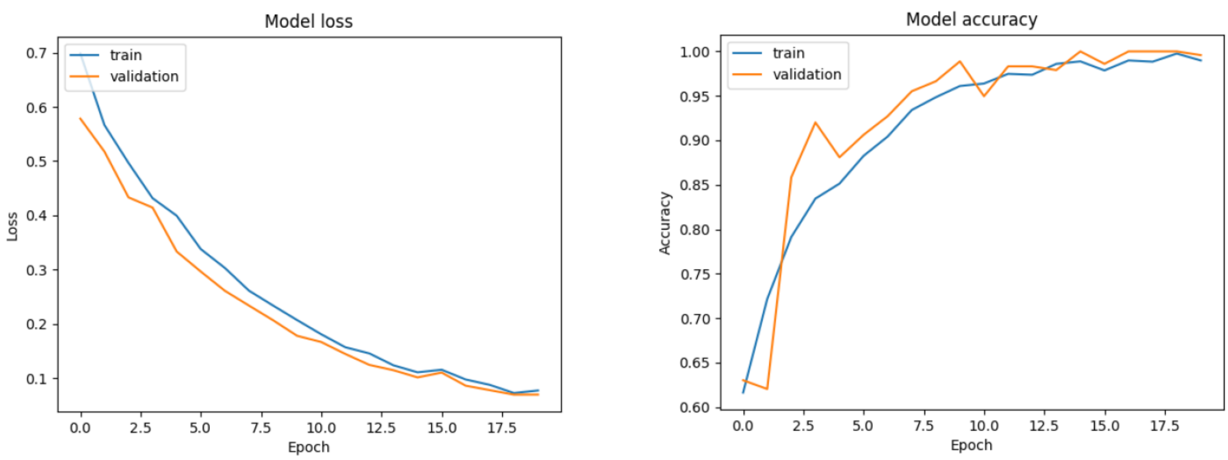


Figure 23. Train/Val loss and accuracy of the hybrid classifier, over 20 epochs

SlowFast

The SlowFast algorithm is based on 3-dimensional well-known convolutional networks ResNets that investigate contrasts of action speed along the time axis of frames. The algorithm consists of two parallel streams with different framerates: a slow pathway and a fast pathway as Figure 57 presents. These two pathways own some significant different temporal speeds, able to recognise action in a sequence of frames. In particular, the fast pathway was designed to understand fast changing motions but with fewer spatial details. On the other hand, the slow pathway was designed to recognise slow changing motions but with a high level of spatial details. The SlowFast algorithm was configured to process rectangle RGB images of 224 × 224 dimensions. Moreover, Adam optimiser with 0.001 learning rate was selected to minimise the categorical cross-entropy loss function. The SlowFast trained for a large number of epochs, equal to 300 aiming to allow the neural network to learn from scratch since it does not employ pre-trained weights.

Additionally, CONV-LSTM architecture was also investigated. Since the data were formed as a sequence of frames, the action recognition task can be transformed into time-series classification using convolutional and recurrent architectures. Aiming to process both the spatial and temporal knowledge, Convolutional and LSTM layers were employed. The Convolutions were responsible to learn the spatial information of a frame. Accordingly, the LSTMs learn temporal features holding information from previous frames of a video. This combination of Convolutions and LSTMs took as input the frame sequence dimensions and the number of time steps that the recurrent part of the architecture look back to recognise possible action. The CONV-LSTM architecture was configured to also process rectangle images of 224 × 224 dimensions. Since this architecture is relative to time, 45 sequence frames were defined, allowing the network to look back at these specific time steps. Adam optimiser was selected to minimise the categorical cross-entropy loss function, as well. The CONV-LSTM employed the pre-trained ResNet18 convolutional model for feature extraction. Accordingly, the model trained for 100 epochs assuming that the neural network will properly converge and generalise.

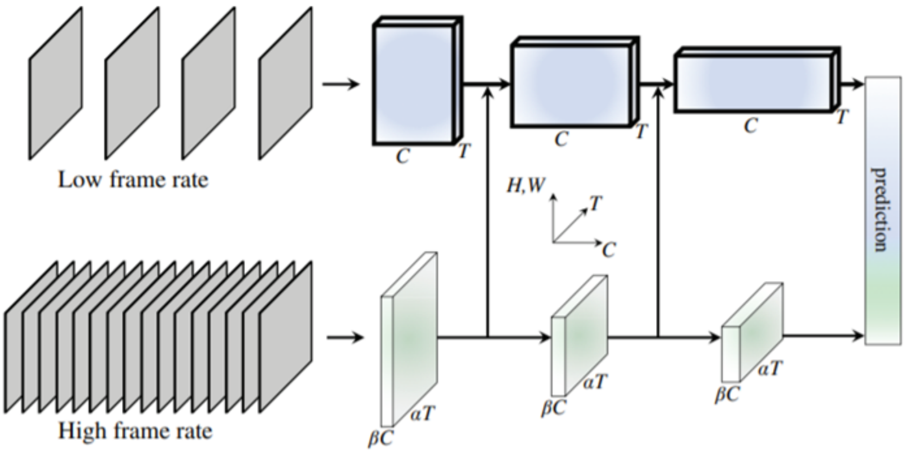


Figure 24. SlowFast network architecture

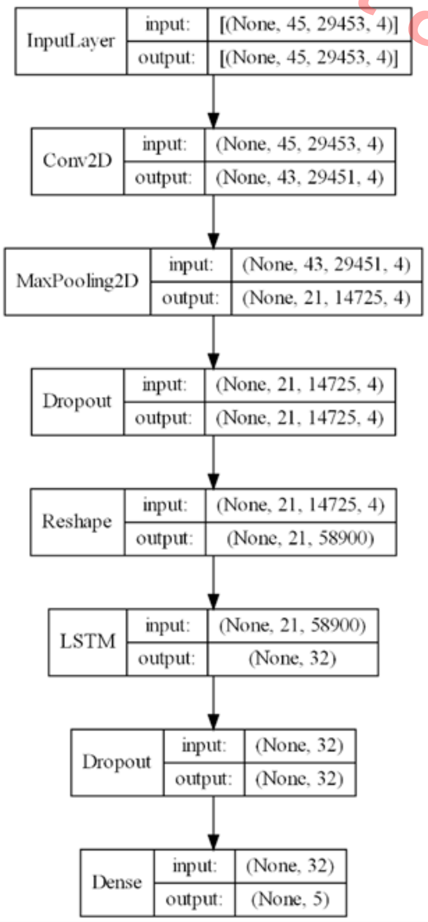


Figure 25. CRNN architecture

3.3.1.2 Sound analysis

This Section focuses on the services that are going to be deployed in the autonomous bus, using information from acoustic sensors. Specifically, we describe the datasets that have been used and the algorithms for the audio-based abnormal event detection starting from the pre-processing steps to the classification deep neural network module.

In sound analysis, it is important to analyse the signal either at its raw form (waveform in time-domain) or convert the input signal to the frequency domain and extract audio features (e.g., spectrogram representation, mel-frequency cepstral coefficients, spectral roll-off, etc.). An advantage of the latter is that the dimensionality and complexity of the data set is significantly reduced, but there is some information lost in the process. In this Section, we will first introduce the MIVIA Audio Events Dataset, and then look at the sounds from a time perspective, frequency perspective and see how to extract information. Finally, we look at some of the samplings and see if we can classify them from a human visual perspective.

Background

In everyday life, we are continuously surrounded by sounds, and usually the human brain is able to recognize what kind of sound it is based on experience. Artificial neural networks are again based on studies of the human brain, and if the artificial neurons work as the biological ones, there should be possible to train a neural network recognizing sounds as well. Lately, immense efforts have been put into this subject with the purpose of translating voice into text, ledby technology companies like Google, Microsoft, Amazon and so on; who develop voice controlled virtual assistants. The technology is promising; the time saving using voice commands compared to the keyboard commands is potentially large and has a great market[22].

Artificial neural networks that use an audio signal as an input have received a great research interest in the past ten years and have shown potential in speech recognition and environmental sound classification. Piczak[23] tested a very simple CNN architecture with environmental audio data and achieved accuracies comparable to state-of-the-art classifiers. Cakir et al.[24] used one dimensional (time domain) deep neural networks (DNNs) in polyphonic sound event detection for 61 classes to achieve an accuracy of 63.8%, which was a 19% improvement over a hybrid Hidden Markov Model/Non-negative Matrix Factorization method. Lane et al.[25] created a mobile application capable of performing very accurate speaker diarization and emotion recognition using deep learning. Recently, Wilkinson et al.[26] performed unsupervised separation of environmental noise sources adding artificial Gaussian noise to pre-labeled signals and used auto-encoders to cluster. However, background noise in an environmental signal is usually non-Gaussian, making this method work on specific datasets only.

Regarding the field of surveillance using audio sensors, there has been extensive research in order to robustly identify the audio scene where an event takes place. The main reason is that audio sensors (microphones) offer the following advantages:

- microphones are cheap, compared to the cameras and can be easily deployed in any kind of environment
- omnidirectional coverage
- specular reflections of the audio signal can be another form of audio input[27]

However, the main challenge that this research area is facing is that environmental sounds are unstructured, especially compared with the speech signals where a phoneme-based approach can be used for classification. The signal-to-noise ratio (SNR) is typically small for an environmental signal, especially when the device is placed far away from the acoustic source. Additional challenges include the

[22] Jiang, Jiepu, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. "Automatic online evaluation of intelligent assistants." In *Proceedings of the 24th International Conference on World Wide Web*, pp. 506-516. International World Wide Web Conferences Steering Committee, 2015.

[23] Piczak, Karol J. "Environmental sound classification with convolutional neural networks." In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1-6. IEEE, 2015.

[24] Cakir, Emre, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. "Polyphonic sound event detection using multi label deep neural networks." In 2015 international joint conference on neural networks (IJCNN), pp. 1-7. IEEE, 2015.

[25] Lane, Nicholas D., Petko Georgiev, and Lorena Qendro. "DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning." In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 283-294. ACM, 2015.

[26] Wilkinson, Bryan, Charlotte Ellison, Edward T. Nykaza, Arnold P. Boedihardjo, Anton Netchaev, Zhiguang Wang, Steven L. Bunkley, Tim Oates, and Matthew G. Blevins. "Deep learning for unsupervised separation of environmental noise sources." The Journal of the Acoustical Society of America 141, no. 5 (2017): 3964-3964.

[27] Baum, Elizabeth, Mario Harper, Ryan Alicea, and Camilo Ordonez. "Sound identification for fire-fighting mobile robots." In 2018 Second IEEE International Conference on Robotic Computing (IRC), pp. 79-86. IEEE, 2018.

recognition of overlapping events from one environment[28], audio event recognition using weakly labelled data[29] and lack of public datasets that contain information from multiple sensors[30].

In this report, we focus on the MIVIA audio events dataset[31], where sounds such as background noise, glass break, gun shot and scream are differentiated. After providing information about the dataset, we see how we can analyze the audio samples in the time and frequency domain and provide details about audio feature extraction. In the Classification Methods Section, we mainly focus on two-dimensional Convolutional Neural Networks (CNNs) that have shown great performance in recognition accuracy[32]. The focus of the work was to study the ability of 2D CNNs to generalize in different SNR conditions. For this work, the Adam[33] optimizer was used and the ReLU[34] activation function between the convolutional layers. Finally, all the accuracy scores are presented in the Results Section.

MIVIA Audio Events Dataset

The MIVIA audio events data set is composed of 6,000 events for surveillance applications, namely glass breaking, gun shots and screams. The 6,000 events are divided into a training set (composed of 4,200 events) and a test set (composed of 1,800 events).

In audio surveillance applications, the events of interest (for instance a scream) can occur at different distances from the microphone that correspond to different levels of the signal-to-noise ratio. Moreover, in these applications the events are generally mixed with a complex background, usually composed of several types of different sounds depending on the specific environments both indoor and outdoor (household appliances, cheering of crowds, talking people, traffic jam, passing cars or motorbikes etc.).

The data set is designed to provide each audio event at eight different values of signal-to-noise ratio (namely -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, 25 dB and 30 dB) and over-imposed to different combinations of environmental sounds in order to simulate their occurrence in different ambiences.

The sounds have been registered with an Axis P8221 Audio Module and an Axis T83 omnidirectional microphone for audio surveillance applications sampled at 32,000 Hz and quantized at 16 bits per PCM sample. The audio clips are distributed as WAV files. The training set has a duration of about 20 hours while the test set of about 9 hours.

The events of interest are organized in three classes (glass breaking, gun shots and screams) and their duration in the training and test sets is reported in the following table

[28] Dikmen, Onur, and Annamaria Mesaros. "Sound event detection using non-negative dictionaries learned from annotated overlapping events." In 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1-4. IEEE, 2013.

[29] Kumar, Anurag, and Bhiksha Raj. "Audio event detection using weakly labeled data." In Proceedings of the 24th ACM international conference on Multimedia, pp. 1038-1047. ACM, 2016.

[30] Chachada, Sachin, and C-C. Jay Kuo. "Environmental sound recognition: A survey." APSIPA Transactions on Signal and Information Processing 3 (2014).

[31] Foggia, Pasquale, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. "Reliable detection of audio events in highly noisy environments." Pattern Recognition Letters 65 (2015): 22-28.

[32] Zhao, Jianfeng, Xia Mao, and Lijiang Chen. "Speech emotion recognition using deep 1D & 2D CNN LSTM networks." Biomedical Signal Processing and Control 47 (2019): 312-323.

[33] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[34] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." In Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814. 2010.

Table 4. The MIVIA audio events dataset

	TRAINING SET		***TEST SET***	
	#Events	*Duration (s)*	*#Events*	*Duration (s)*
Background	10768	58371,6	4716	25036,8
Glass breaking	4200	6024,8	1800	2561,7
Gun shots	4200	1883,6	1800	743,5
Screams	4200	5488,8	1800	2445,4

In our experiments, we have focused on the 0 dB SNR, which is the second most challenging SNR value. At 0 dB the background noise can significantly mask the target sound event, such as the scream. Finally, we notice that the dataset is quite imbalanced towards the background noise. Therefore, our system needed to be robust against the imbalances and the error was evaluated with the F1 macro average score defined as

$$F1_{macro} = \frac{2 \times Precision_{macro} \times Recall_{macro}}{Precision_{macro} + Recall_{macro}} \qquad \textbf{(1)}$$

By macro averaging, we compute the metric independently for each class and then take the average, hence treating all classes equally.

Time Domain: Sounds are longitudinal waves, which actually are different pressures in the air. They are usually represented as a function in the time domain, which means how the pressure varies per time. This function is obviously continuous, but since computers represent functions as arrays, we cannot save all the information.
When discretizing the audio signal, we lose some information. The loss of the information depends on the sampling rate, which is the number of sampling points per second (Figure 18). According to the Nyquist theorem, one should have twice as high sampling rate as the highest sound frequency to keep most of the information and avoid clipping of the signal. For instance, a human ear can perceive frequencies in the range 20-20,000Hz, so a sampling rate around 40 kHz should be sufficient to keep all the information. Ordinary CD's use a sampling rate of 44.1 kHz, but for speech recognition, 16 kHz is enough to cover the frequency range of human speech.

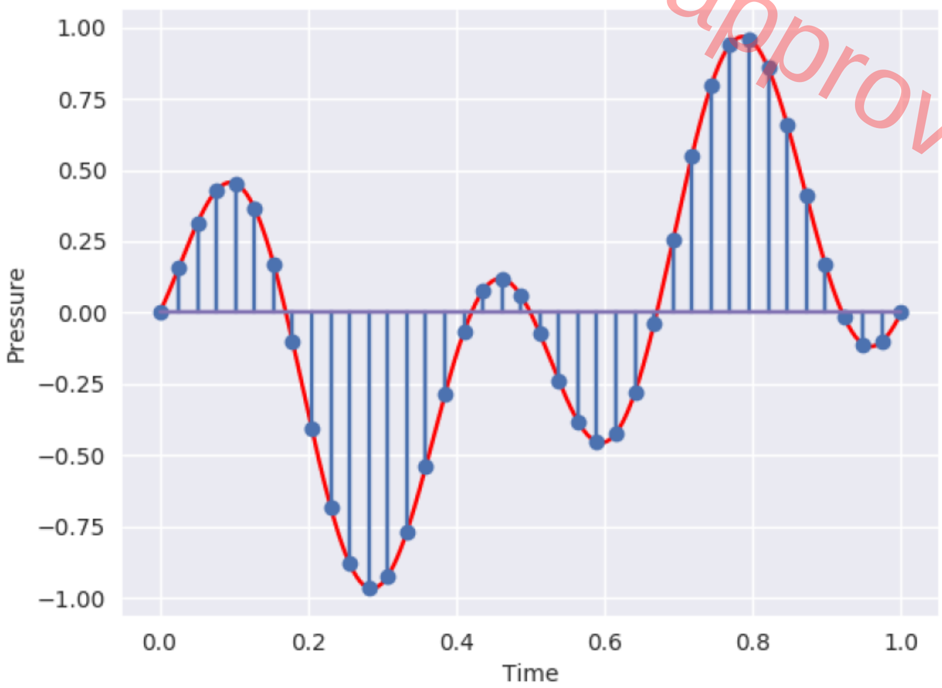


Figure 26. Quantized audio signal (continuous to discrete)

Frequency Domain: Most of the time, a frequency representation of the signal provides a better picture than the time domain. On one hand, one loses information in the time domain, but on the other one, gets information about which frequencies are significant in the original audio signal. To transform from the time domain to the frequency domain, the Fourier Transformation is performed, defined by

$$\hat{f}(x) = \int_{-\infty}^{\infty} f(t)e^{-2\pi ixt}dx \qquad \textbf{(2)}$$

In the above equation, $f(t)$ is the time function and $\hat{f}(x)$ is the frequency function. The Fast Fourier Transform (FFT) has been proposed in order to achieve higher speeds in calculating the Fourier Transformations by splitting the input signal into many sub-signals. In Figure 20, the FFT is calculated on the time domain in order to obtain the corresponding frequency domain.

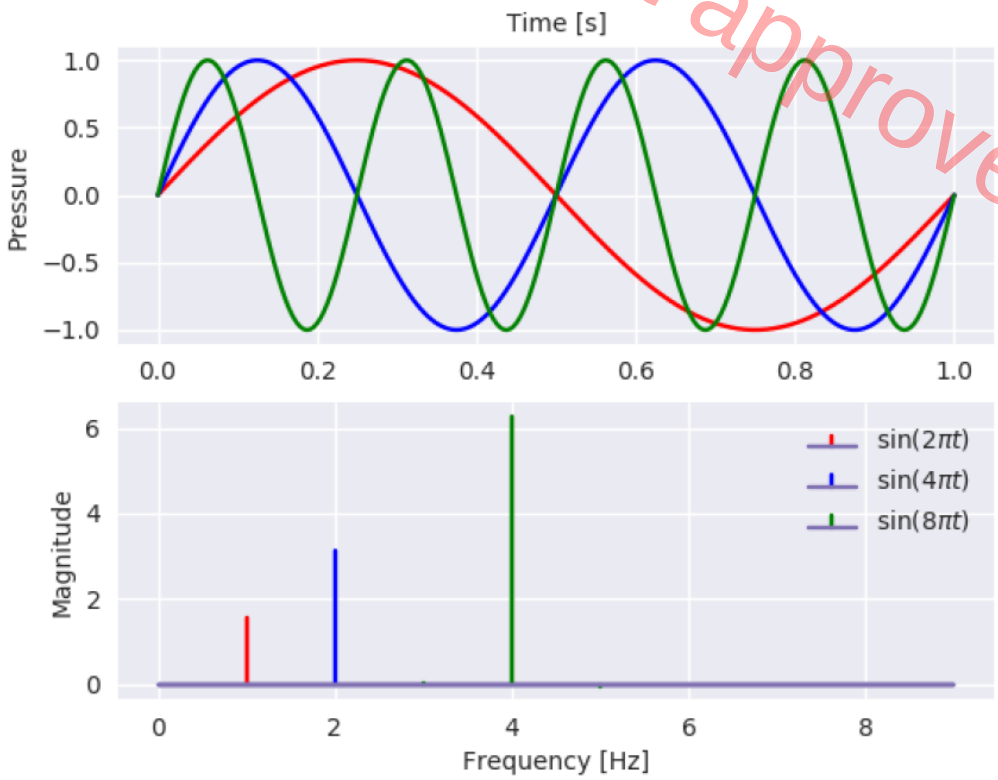
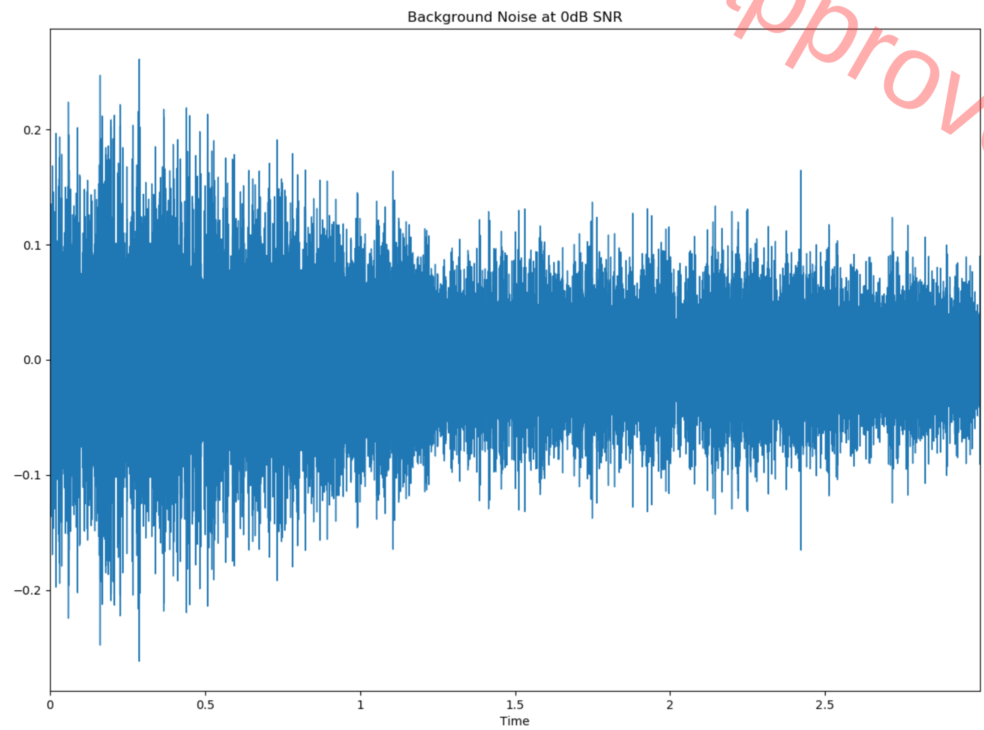
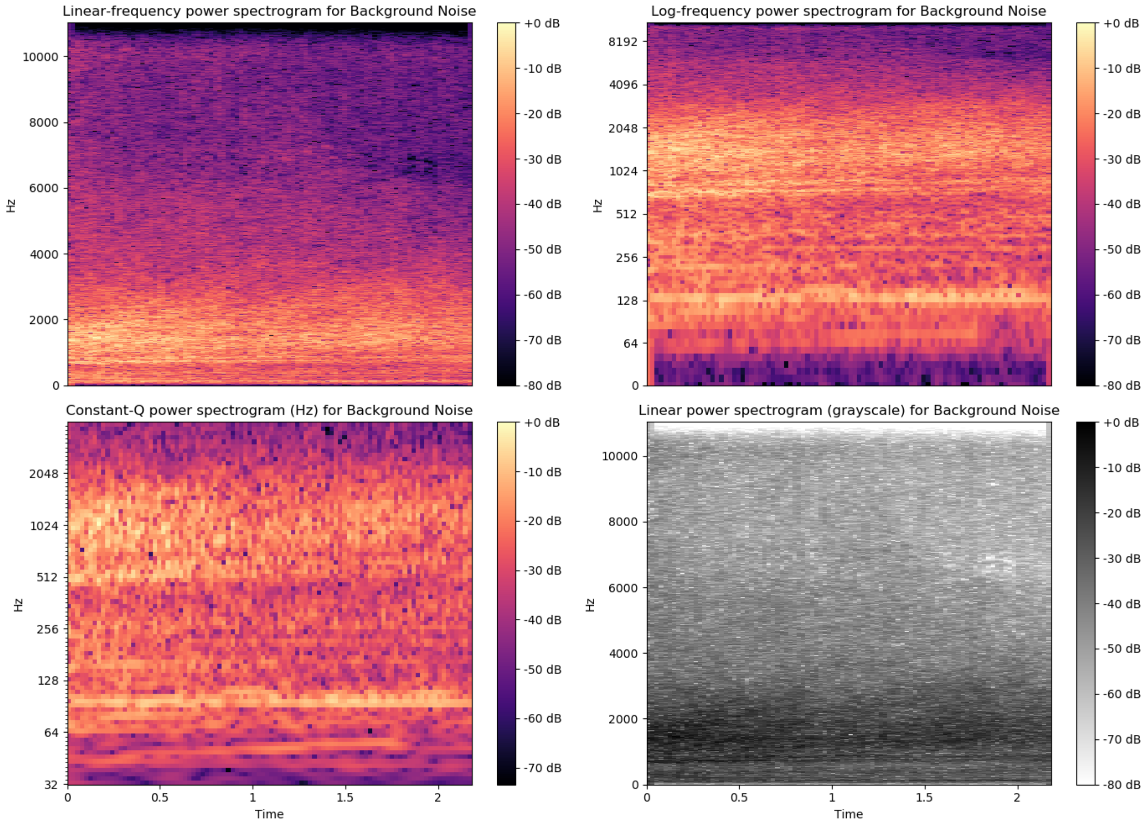


Figure 27. Waveforms of three harmonic functions (top) and their corresponding frequency responses (bottom)

Feature Extraction: To include the range of frequencies that are relevant to identifying the given environmental sounds and to efficiently extract the audio features, we split the input signal into smaller frames for processing. First, we down-sampled the original 32 kHz audio signal to 16 kHz that allowed faster processing. Each frame had an FFT window size of 512 with a 256 hop length (50% overlap). Therefore, we resulted in a 188 × 257 grayscale spectrogram image, for a 3 s recording, that was used as an input to the 2D CNN classifier.

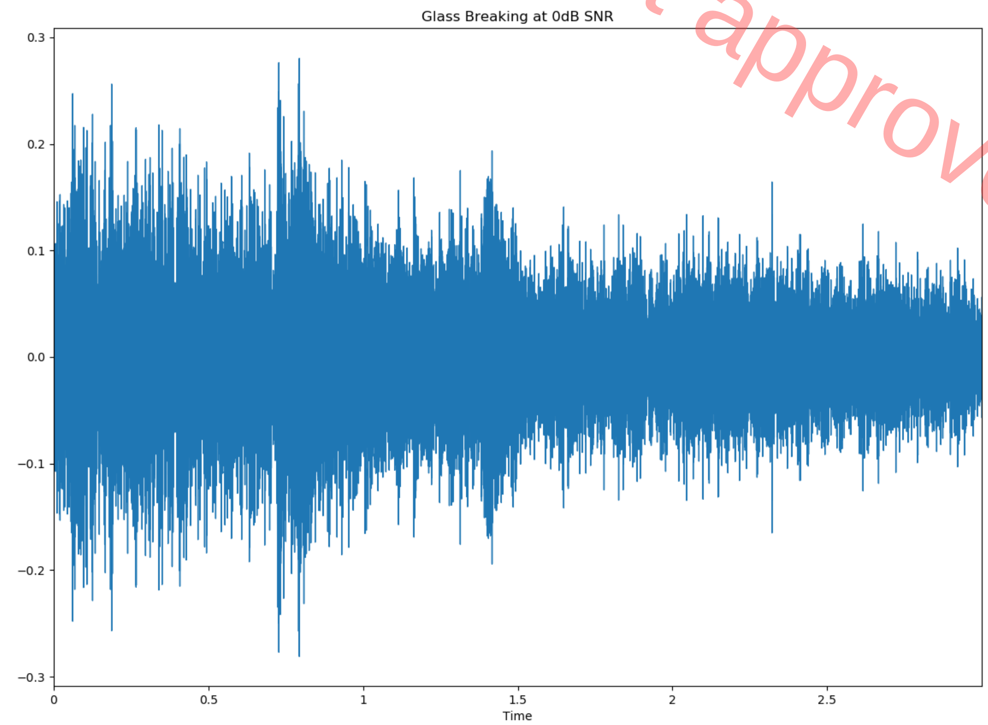


(a)

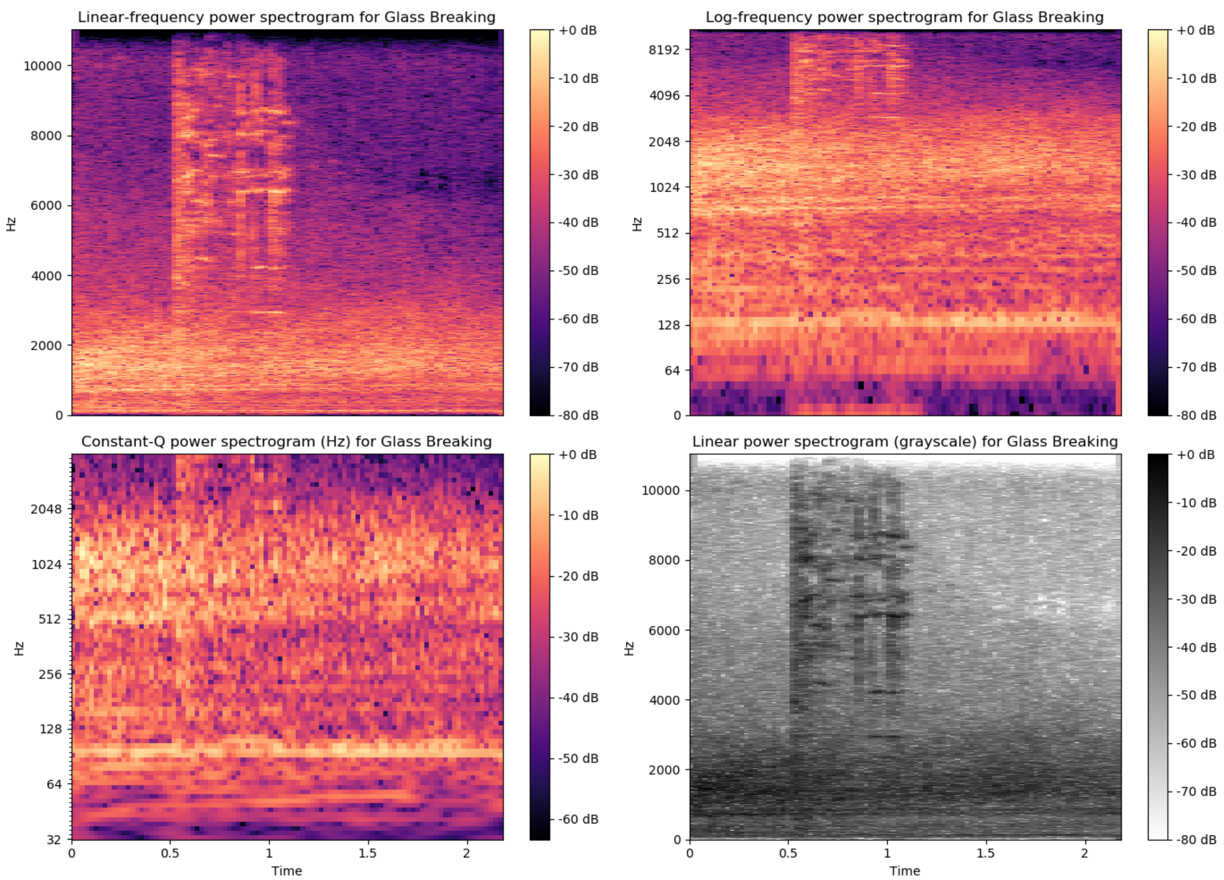


(b)

Figure 28. Time (a) and frequency (b) representations of the background noise at 0 dB SNR

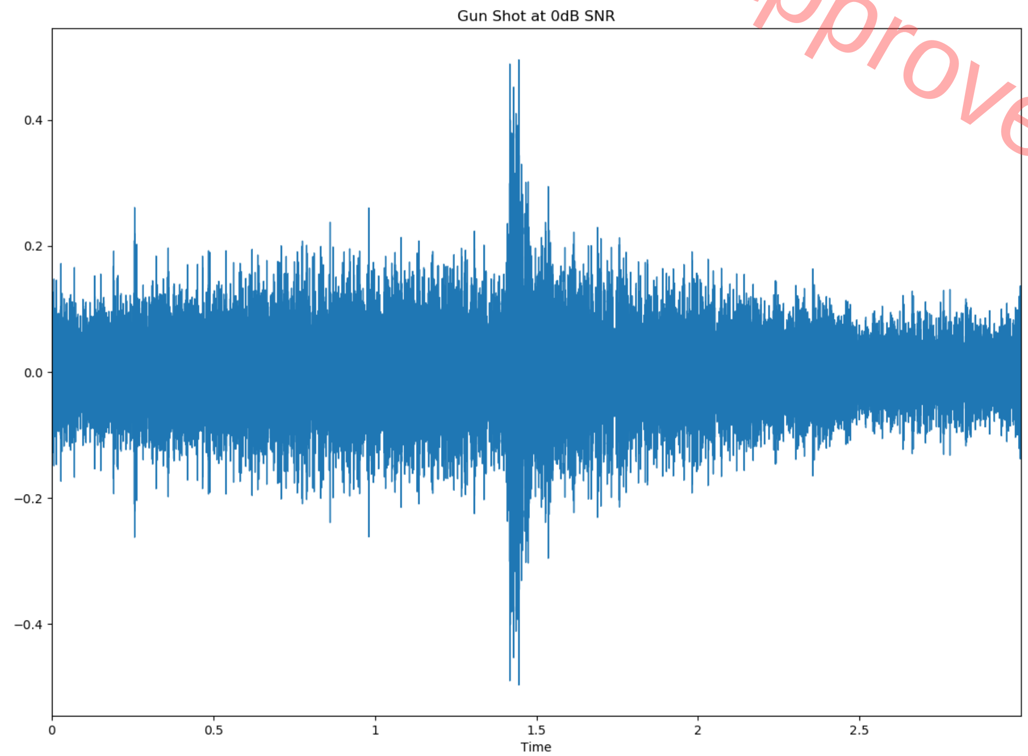


(a)

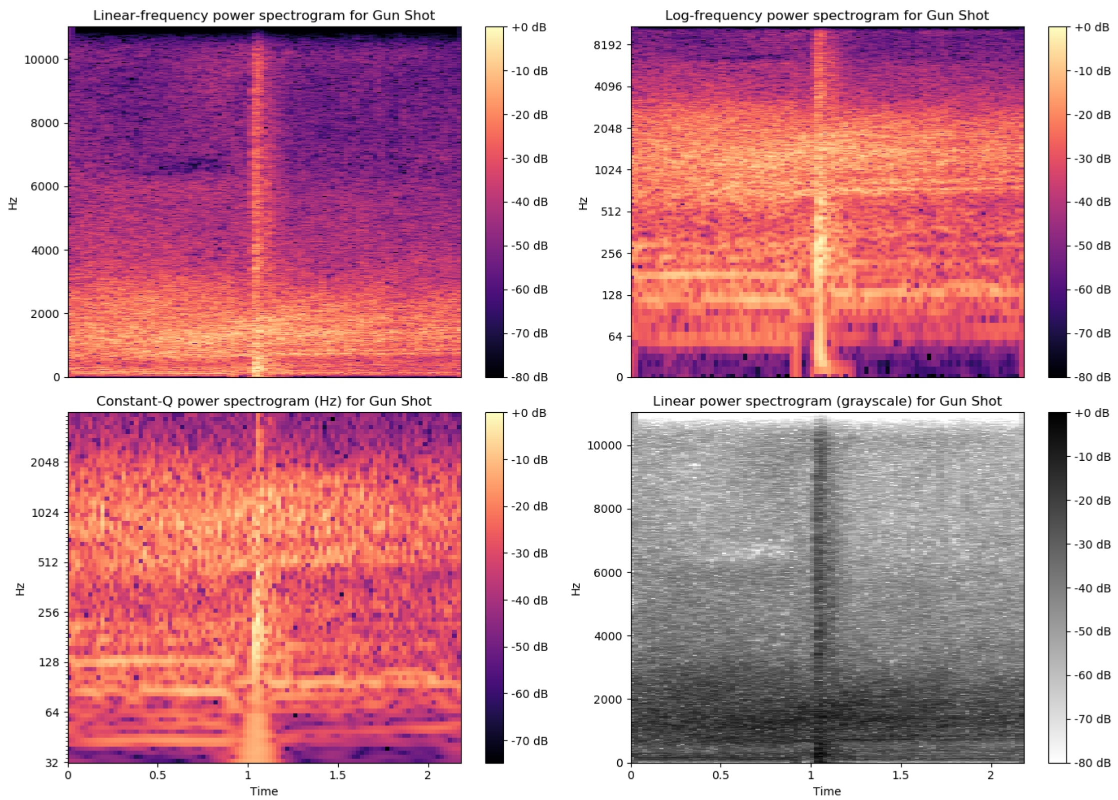


(b)

Figure 29. Time (a) and frequency (b) representations of a glass breaking at 0 dB SNR

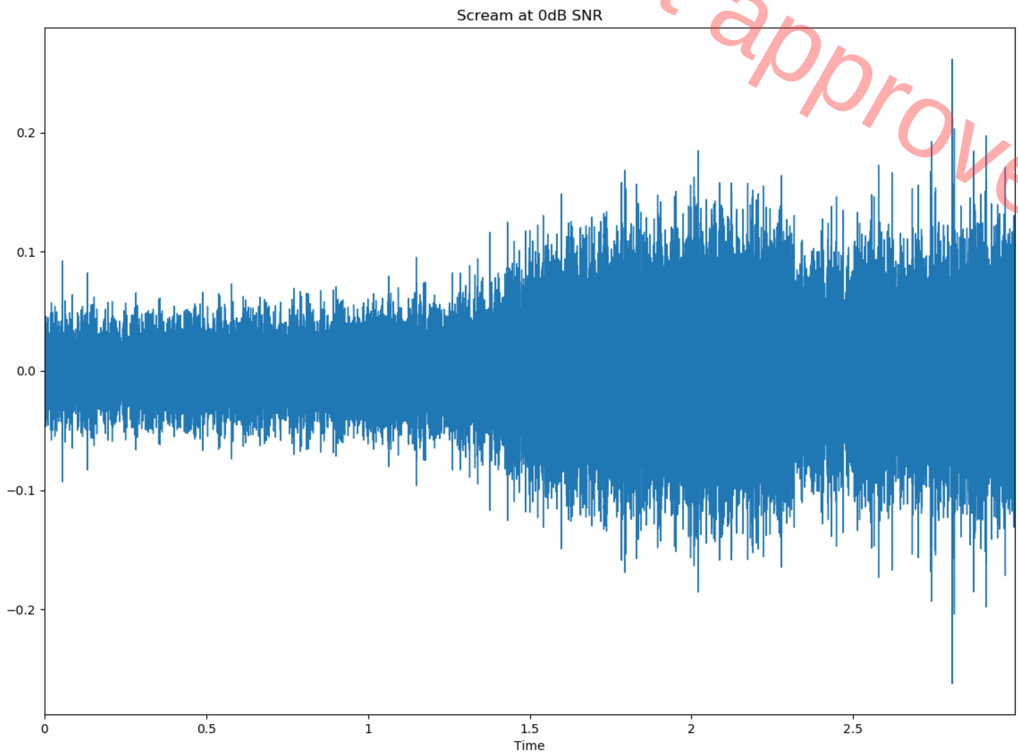


(a)

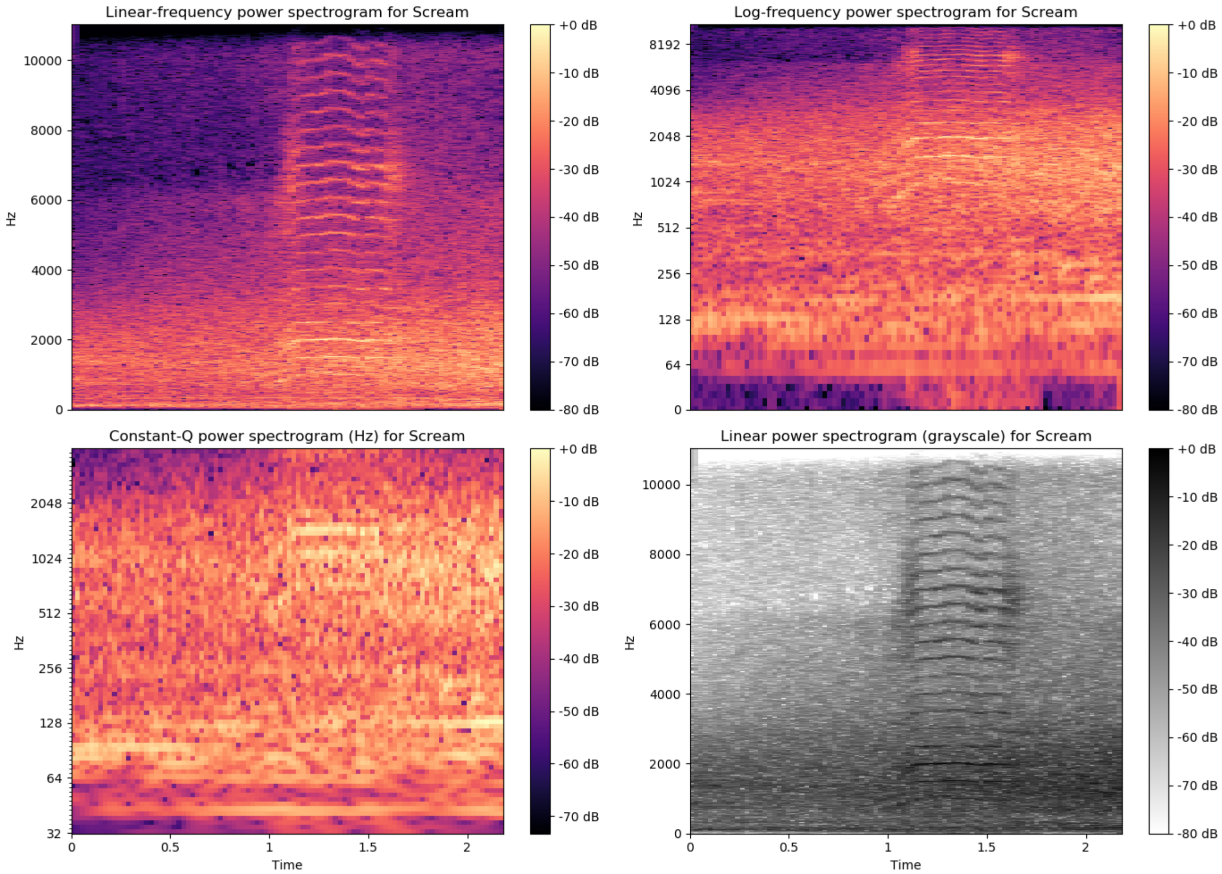


(b)

Figure 30. Time (a) and frequency (b) representations of a gunshot at 0 dB SNR



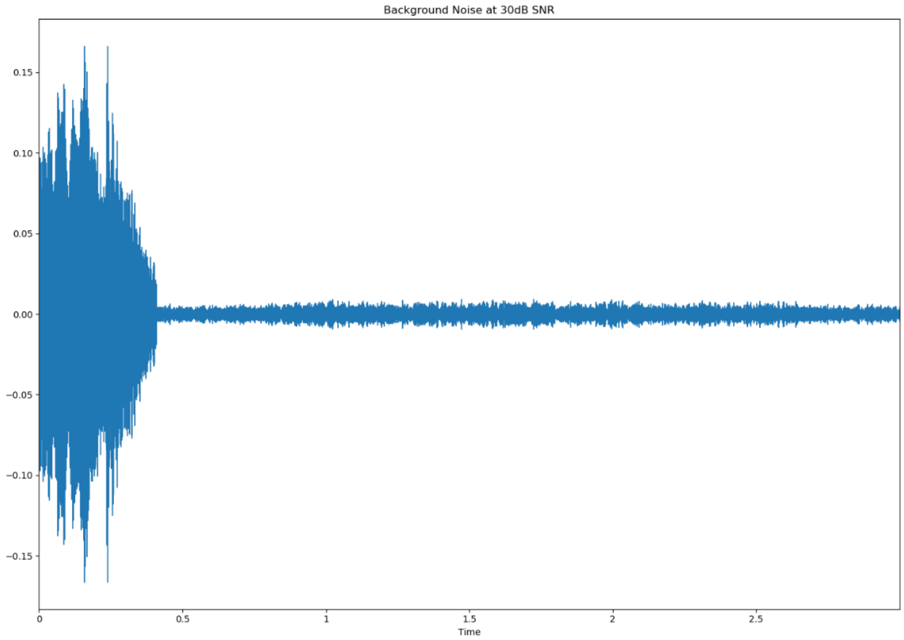
(a)



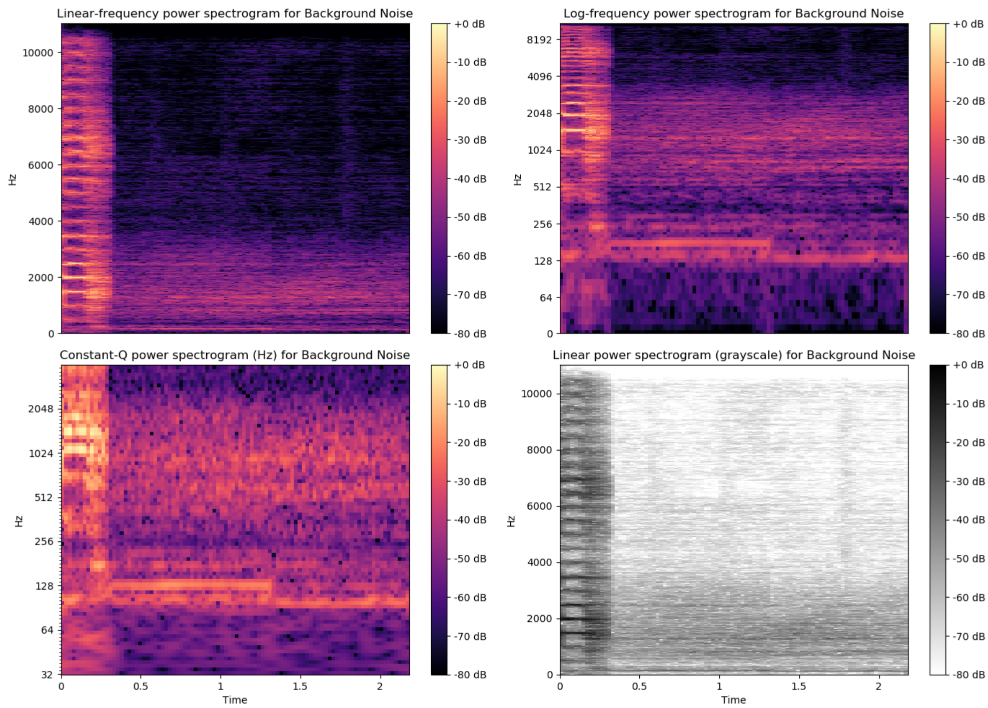
(b)

Figure 31. Time (a) and frequency (b) representations of a person screaming at 0 dB SNR

Figures 28-31 show the time and frequency domain representations of the selected target classes at 0 dB SNR (noisy signal). One can notice that the waveform of the background noise and the glass breaking look very similar in a very noisy environment. However, when observing the FFT spectrogram, the differences are clearer, since there are some high energies concentrated in frequencies between 4 and 10 kHz. Additionally, in this work, we have studied the audio signals with SNR values varying between -5 dB and 30 dB with 5 dB increments in between. Figures 32-35 show the time and frequency domain representations of the selected target classes at 30 dB SNR (clean signal). We notice that the signal at 30 dB SNR is much cleaner, compared to the one at 0 dB and therefore, the network would be able to classify the target classes with higher accuracy.

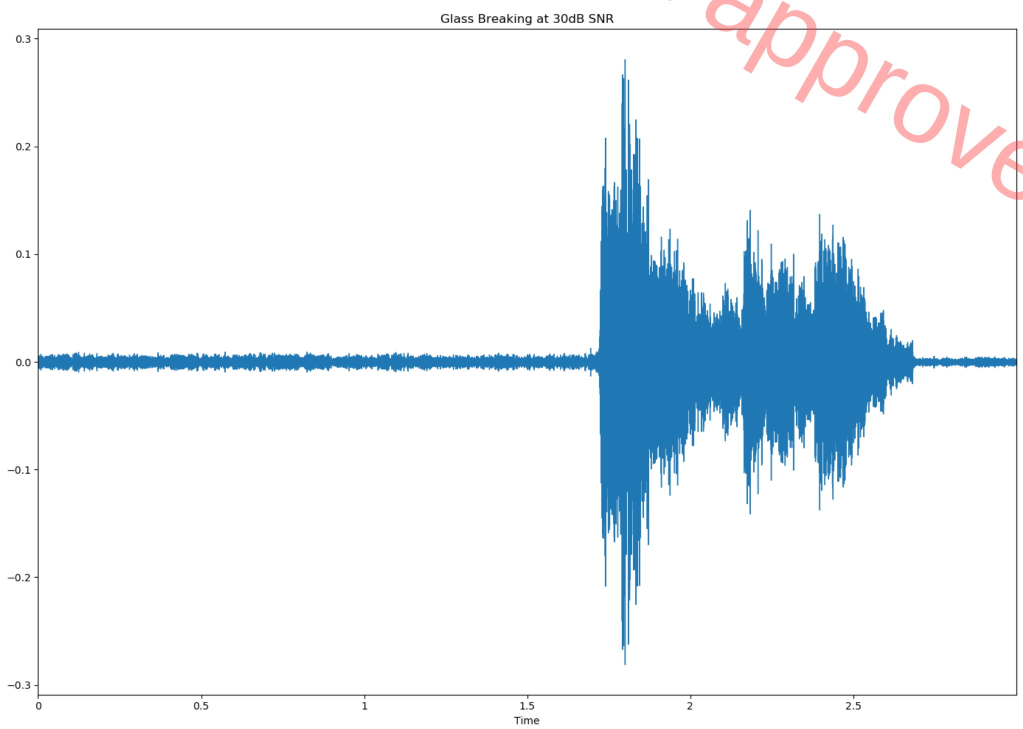


(a)

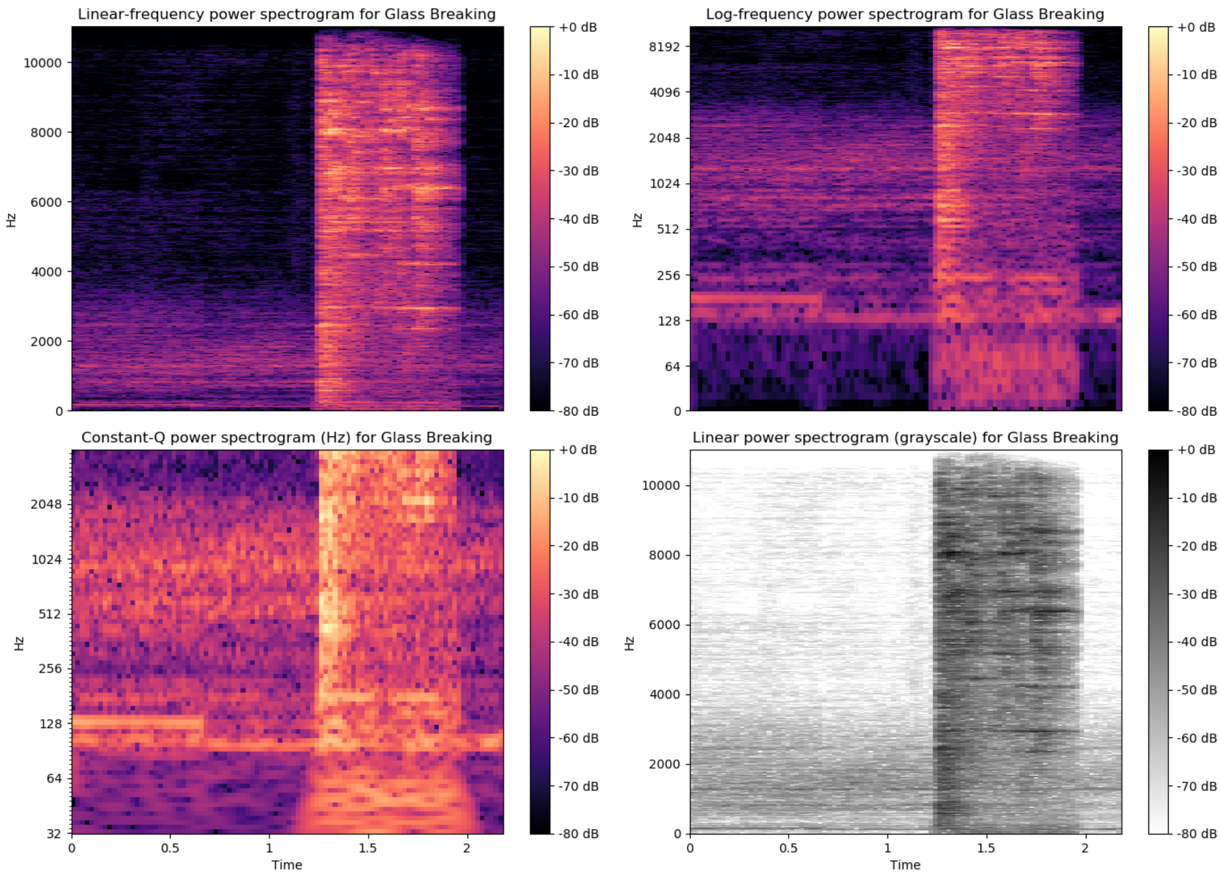


(b)

Figure 32. Time (a) and frequency (b) representations of the background noise at 30 dB SNR

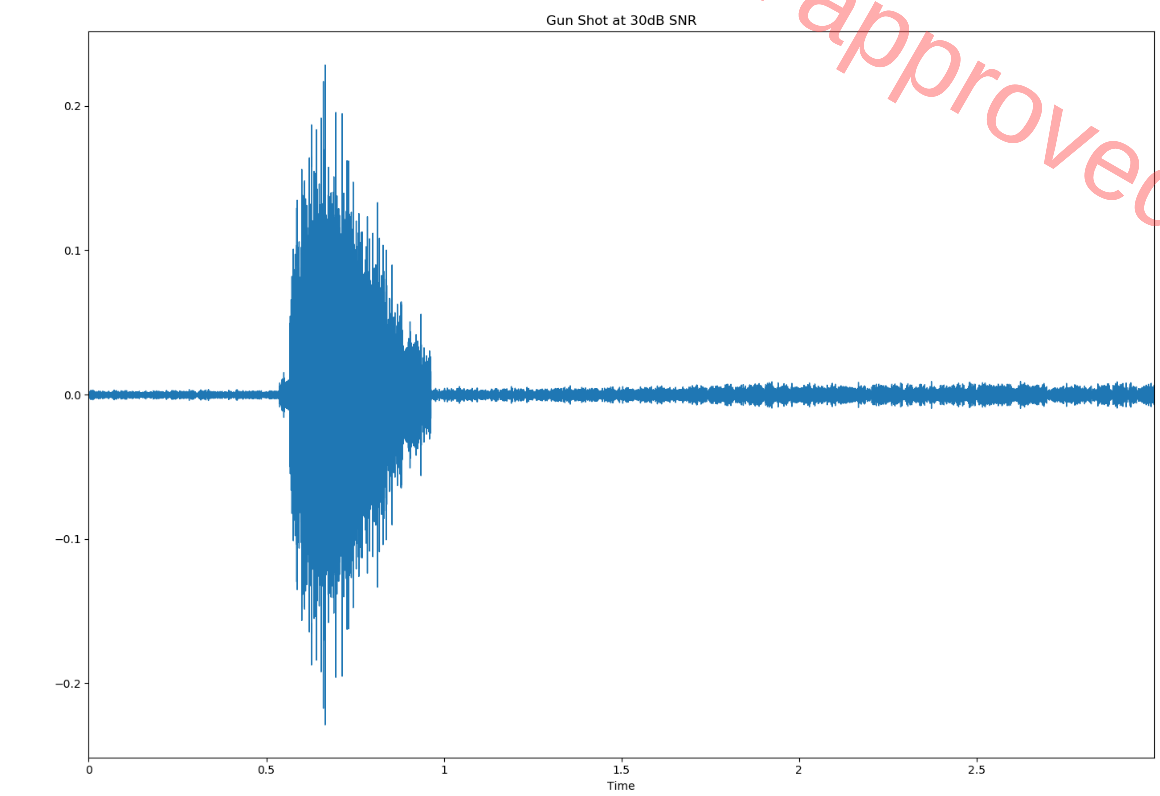


(a)

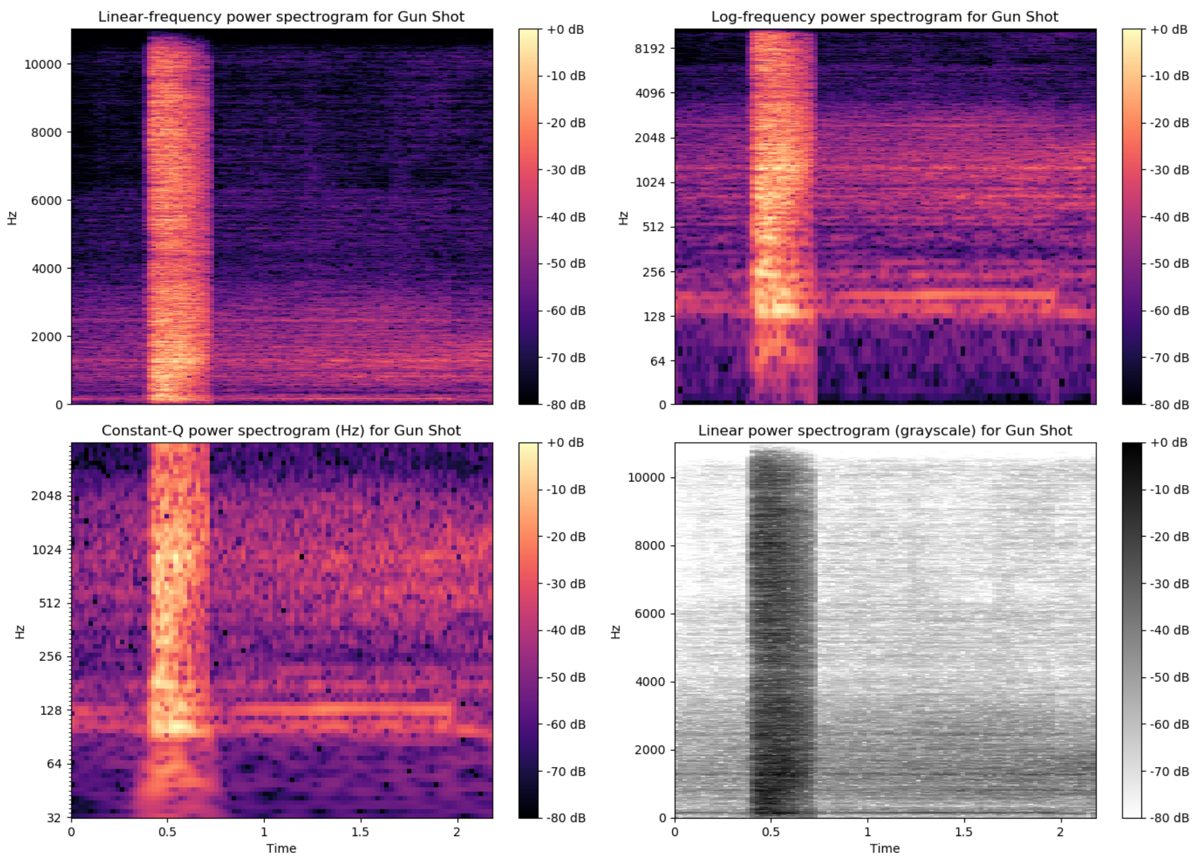


(b)

Figure 33. Time (a) and frequency (b) representations of a glass breaking at 30 dB SNR

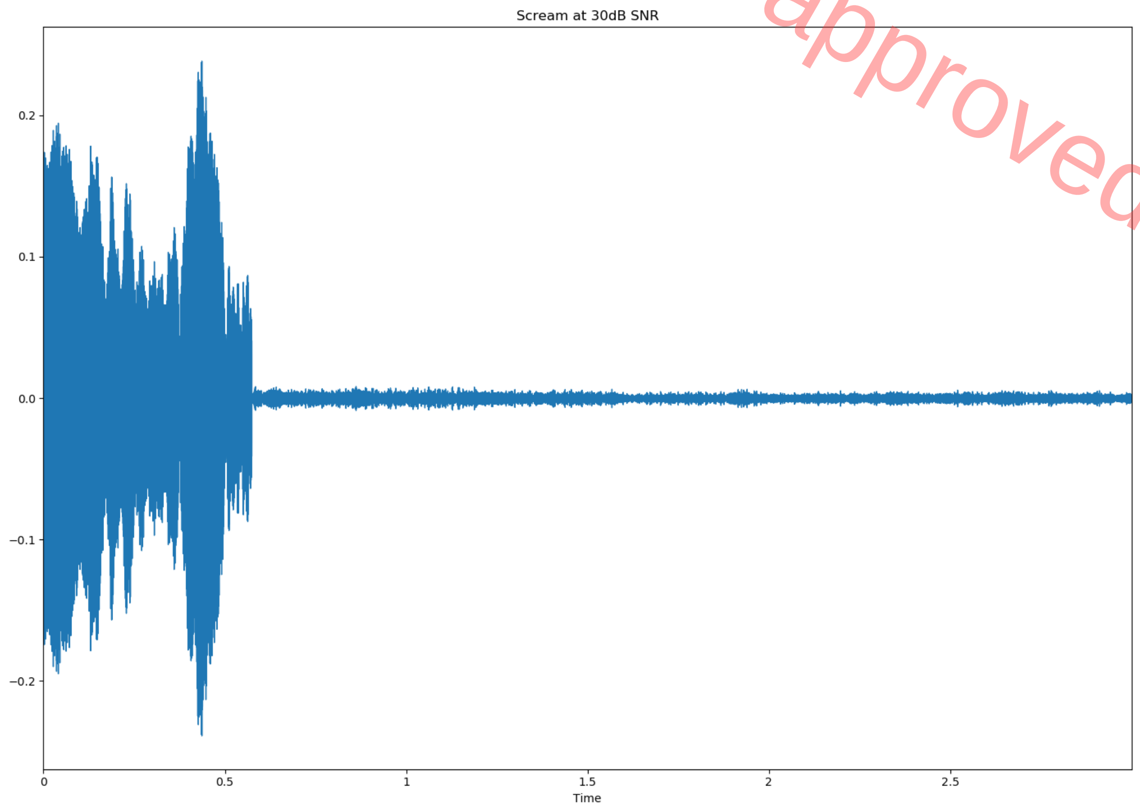


(a)

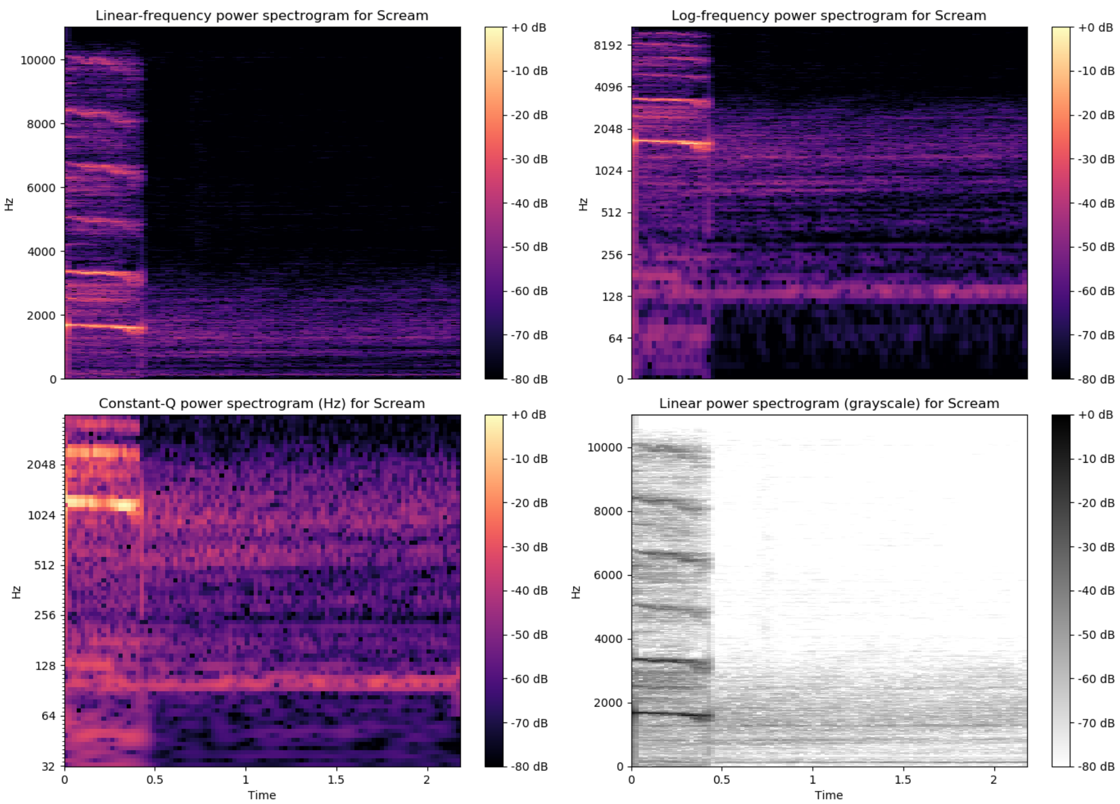


(b)

Figure 34. Time (a) and frequency (b) representations of a gun shot at 30 dB SNR



(a)



(b)

Figure 35. Time (a) and frequency (b) representations of a person screaming at 30 dB SNR

In order to be able to compare the performances of the frequency representations, we considered the case of a so-called “multichannel” representation of the raw audio. In particular, three of the most popular in the literature representations (STFT, mel and MFCC) were combined following two separate methods:

(i) The first was by concatenating the 3 channel (RGB) representations of each spectrogram feature together. That would mean simply adding the (188 × 257 × 3) STFT, the (188 × 128 × 3) mel and the (188 × 60 × 3) MFCC spectrogram frequency features, resulting in a multichannel (188 × 445 × 3) sum of features to be used as input for the model with regard to the concatenated method.

(ii) The second method was to stack the different spectrograms in their grayscale form together, as different channels. In order for that to happen, each spectrogram was reshaped to a common feature-time dimension length, which was chosen to be 224 × 224 (see below). To that end, an area interpolation algorithm was used (resampling using pixel area relation). It is a commonly used method for image decimation, as it is reported to provide moiré effect-free results, while in the case of image interpolation, it is known to yield similar results to the nearest neighbour interpolation method. The resulting shape of each representation derived from this process was 224 × 224 × 3. Then, the grayscale spectrogram of each representation was used as per Equation (2). Finally, with each representation being 224 × 224 × 1, they were all stacked together, comprising a 224 × 224 × 3 representation (the TensorFlow preprocessing *array_to_img* method was used to produce the final image from the array) with each channel being either the STFT, mel or MFCC spectrogram instead of the RGB channels. These would be the dimensions of the input representation for the model with regard to the stacked method.

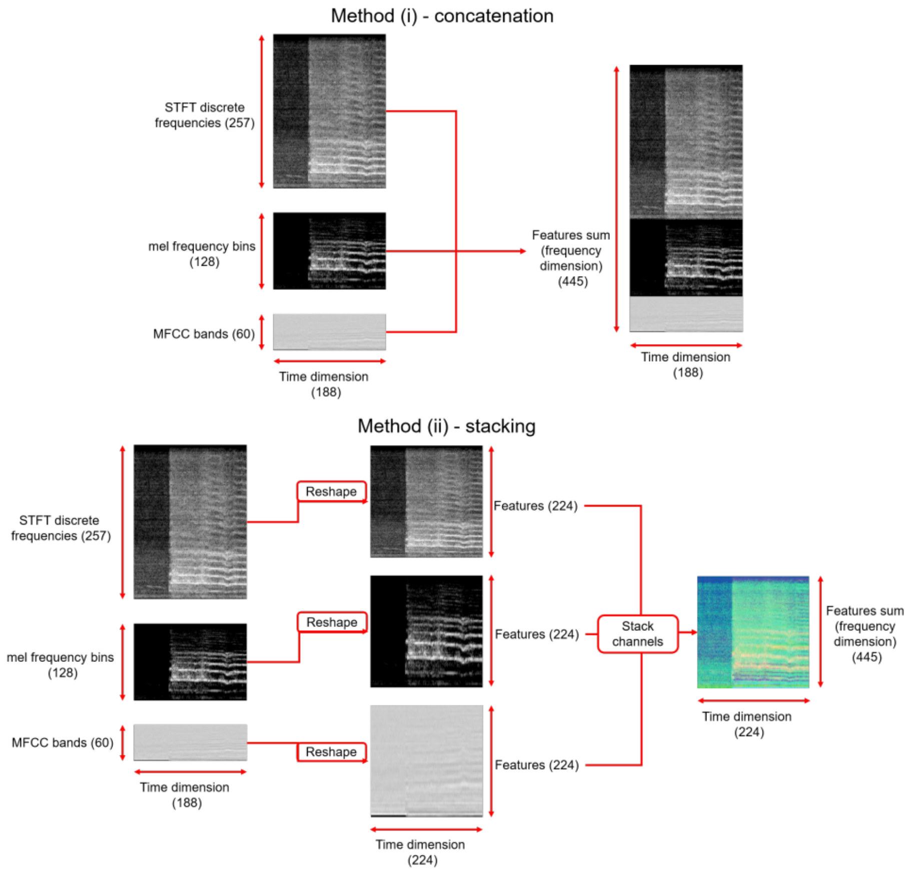


Figure 36. The two representation methods: the concatenation method is shown on top, and the stacking method is on the bottom

Classification Methods

In this Section, we will describe the Convolutional Neural Network architectures used, the optimizer that was selected and the activation function between the convolutions and max-pooling operations. CNNs are known to perform well in image classification and have achieved similar results in the field of audio-based event detection. In general, the input image is passed through various convolutional layers, before it is flattened and fed to a fully connected neural network that outputs the probabilities of the target classes.

Convolutional Layer: Initially, the image is passed through a convolutional layer, which is meant to reveal the structures and shapes in the image. The way that this is performed is to introduce a filter that slides over the entire image and multiplies with all the pixels (with overlap). Every time the filter is multiplied with a set of pixels, we sum all the multiplications and add the value to an activation map (convolution operation). The activation map is completed after the filter is multiplied with the entire image. A typical filter has dimensions of 16×16 or 32×32, depending on the shape of the input image. It is important in this case, to choose a filter that is large enough to cover all the structures of the image.

Pooling Layer: A pooling layer is a way to reduce dimensionality of the representation (down-sampling) such that there are not many parameters that need optimization, but it also helps to control overfitting. The activation map is divided into regions of equal size and represents each region with one single number. Max-pooling is one of the most popular pooling techniques, which just represents each region with the largest number in that region. However, it is possible to use average pooling, global pooling, etc.

Batch Normalization: To increase the stability of a neural network, batch normalization normalizes the output of a previous activation layer by subtracting the batch mean and dividing by the batch standard deviation. Consequently, batch normalization adds two trainable parameters to each layer, so the normalized output is multiplied by a “standard deviation” parameter (γ) and add a “mean” parameter (β).

Dropout: Dropout is widely used in neural network layers and is another way to prevent overfitting. This layer is simple in the sense that it randomly drops out units (activation maps) in the current layer by setting them to zero.

Fully Connected Layer: The last part of a convolutional network is often referred to as a fully connected layer, which is a regular feed-forward neural network (FNN). The output from the other layers needs to be flattened out before it is passed into the FNN. The activation function that is commonly used for classification is Softmax. This activation function assigns probabilities between the target classes and the probability that is closest to one is selected.

DenseNet-121: As a first part of our experiments, we have used a DenseNet-121 CNN architecture as seen in Figure 37 and Figure 38.

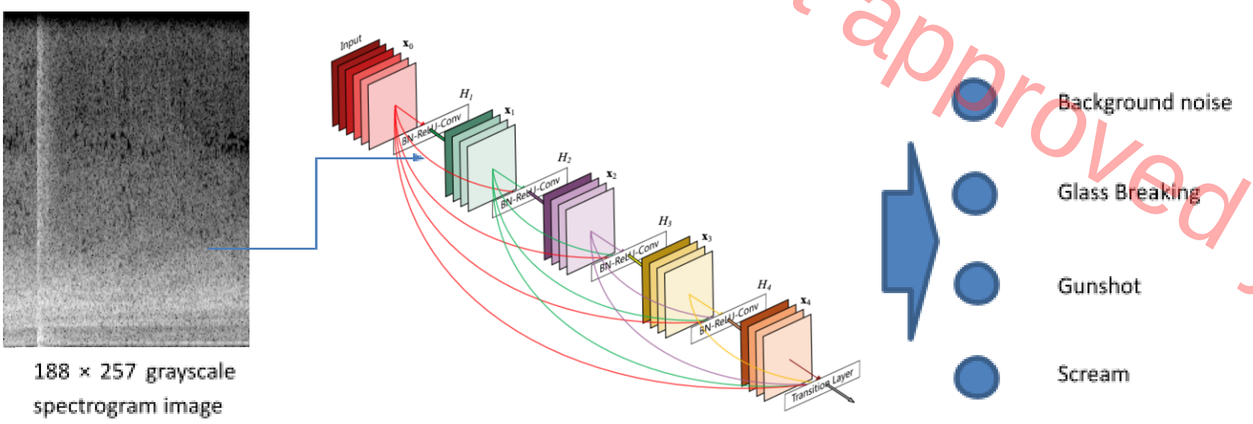


Figure 37. Pipeline of the 2D CNN used for the audio-based event detection

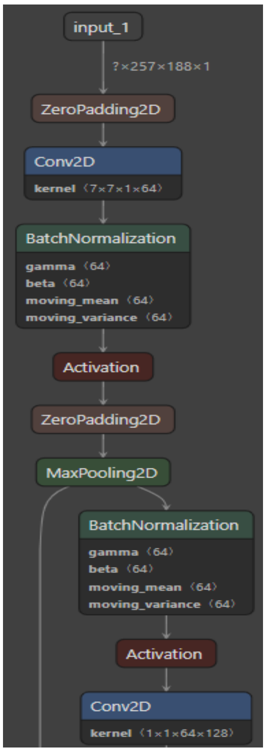


Figure 38. Part of the DenseNet-121 architecture as depicted using the Netron software

DenseNets[35] were introduced in order to solve the problem of the vanishing gradient in neural networks. The vanishing gradient occurs when the CNNs become big that the path for information from the input layer to the output layer significantly increases that the gradient vanishes during back-propagation. In DenseNets, we connect every layer directly with each other. This process ensures maximum information kept throughout the architecture. Additionally, by using this connection the DenseNets require fewer parameters to train than vanilla CNNs, since there is no need to learn redundant feature maps.

The DenseNet-121 was used for classification, as the most basic DenseNet yet powerful architecture. Each dense layer consists of two convolutional operation as follows

[35] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017.

- 1 × 1 Convolutions (for feature extraction)
- 3 × 3 Convolutions (bringing down the feature depth/channel output)

The DenseNet-121 consists of six such dense layers in a dense block. This resulted in approximately seven million parameters, compared to the 44 million parameters of a vanilla CNN architecture.

Optimization

After the non-linear function describing the dataset is found, one typically wants to find the minimum or maximum to optimize various parameters. In order to perform that there are numerical optimization methods. In our experiments, we have used the Adam optimization method with an initial learning rate of 0.001.

Adam is often the preferred optimization method in machine learning, due to its computational efficiency, little memory occupation and implementation ease. Adam is a momentum-based method that only relies on the first derivative of the cost function. Required inputs are exponential decay rates β_1 *and* β_2, cost function c($\boldsymbol{\theta}$) and initial weights $\boldsymbol{\theta}$, in addition to the learning rate η and eventually a regularization λ. Iteratively, the gradient of the cost function is calculated and this is used to calculate the first and second moment estimates.

Activation

The activation function is used to activate the outputs, which often need to take some certain properties. For example, when doing classification, the outputs are probabilities and therefore take values between zero and one. A nonlinear activation function is often preferred to reinforce the non-linearity of the neural network.

Traditionally, the logistic function and the tanh function have been used as activation functions, since they were believed to work in the same way as the human brain. However, in 2012 Alex Krizhevsky et al. introduced AlexNet, taking image recognition to a new dimension. They used a function named Rectified Linear Units (ReLU), which is zero for negative values. The ReLU function is a modification of the pure linear function for positive values. This makes the function able to recognize the non-linearity in the model, providing it a “clean” derivative given by the step function.

3.3.1.2.1 Research results

In the next Section, we will present results from the vision and audio modalities. Future research should consider the potential effects of fusing video and audio modalities. Models can be fused both on decision-level and by concatenating their respective fully connected layers. Recent studies by Kampman et al.[36] and Ortega et al.[37] has proven that using multiple modalities combined allows interaction between them in a non-trivial way and greatly outperforms the individual performance. By combining the last network layers and fine-tuning the parameters, we can take advantage of the complementary information of visual and auditory modalities outperforming the current individual results.

[36] Kampman, Onno, et al. "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018.

[37] Ortega, Juan DS, et al. "Multimodal fusion with deep neural networks for audio-video emotion recognition." arXiv preprint arXiv:1907.03196 (2019).

Video analysis results

We tested our model at the NTU RGB+D dataset and on the data captured by CERTH inside the AV's shuttle (Figure 39). In the images there are various debug layers enabled, such as skeleton points, lines, tracker id and bounding boxes of each detection. The predicted result is marked as green, when the classifier indicates it as "normal" and red when "abnormal", correspondingly. So far, we did not include NTU dataset samples in our training set, so it is safe to assume that our model can generalize across different people, view angles etc. Also in Figures 38-41, we present some results on each use case along with the link to the full video. Classification results and metrics for each class are presented in Table 3.

LSTM Classification results

P: Person Identifier (e.g P1 for the Person with ID1)

Normal behaviour

Abnormal behaviour

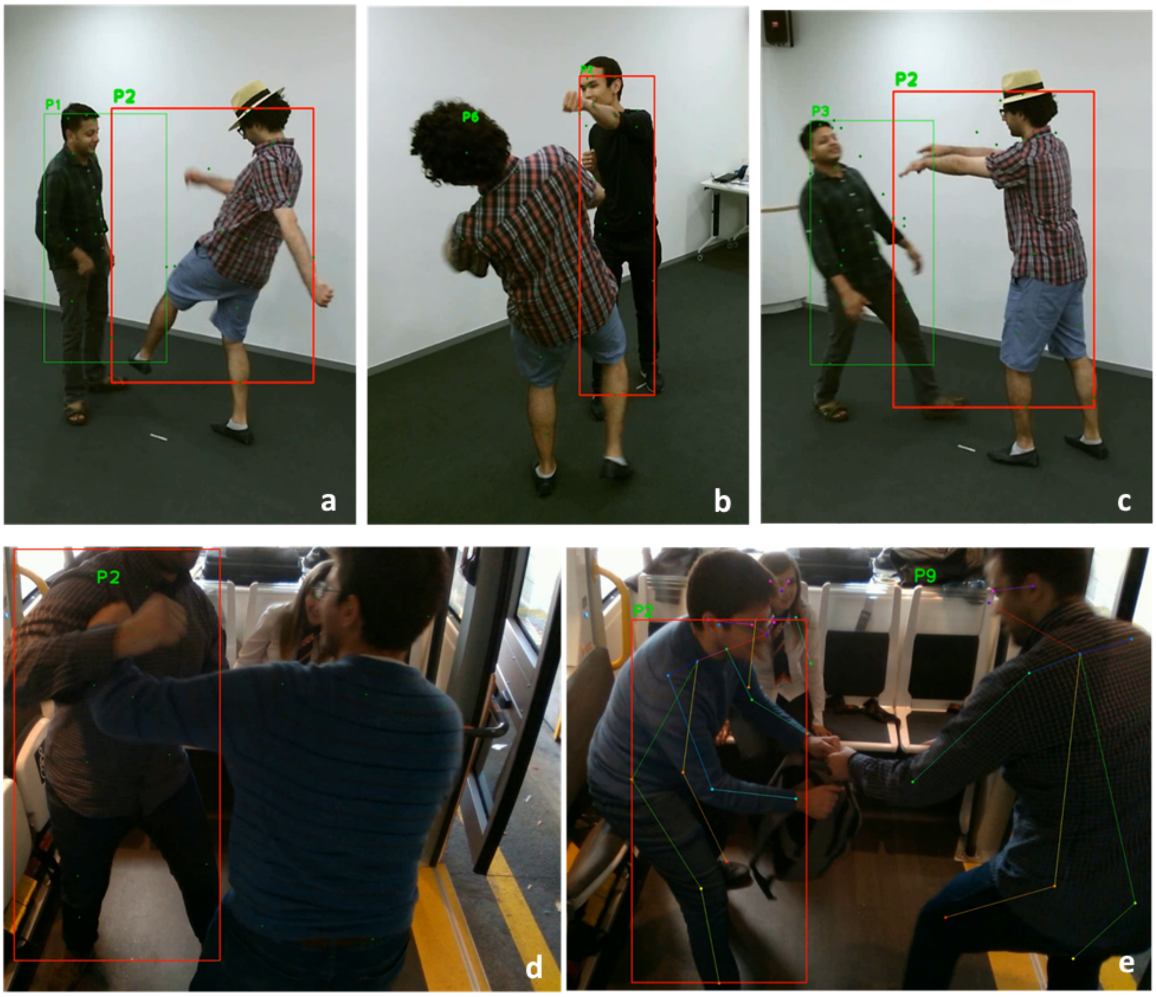


Figure 39. Evaluation on test data:
(a), (b), (c) Abnormal event detection (violence/passengers are fighting) using different camera angles from the NTU-RGB dataset.
(d), (e) Detection of Fighting / Bag Snatch real-world scenarios inside the shuttle.

P: Person Identifier (e.g P1 for the Person with ID1)

Normal behaviour

Abnormal beh

Evaluation on Use Cases

Use Case 2 – Bag Snatch

Figure 40. The commuter is attacked by another person who is attempting to snatch his bag

Full video: https://drive.google.com/open?id=15tLAGcjp5pzI5erHRcg0u2Mh53PSeIkt

Use Case 2 – Fighting 1

Figure 41. A petty crime takes place in the form of assault

Full video: https://drive.google.com/open?id=1fGktDygbtxWT20ILfqBXPs3fdy8mXozi

Use Case 3 – Vandalism

Figure 42. A person attempts to perform a vandalism action on the bus, through painting a graffiti on the windows

Full video: https://drive.google.com/open?id=125B7YX1w_yKqvr6zyiai22urGXK0nV0j

Use Case 1 – Unaccompanied Luggage Monitoring

Figure 43. Unaccompanied luggage which remains as is unmoved for a long time

Full video: https://drive.google.com/open?id=1rCV0nYH4LQwOv4hlw1vyeXpXqcdrByMw

Detection of certain events may raise a notification or an alert to the supervisor and/or the suitable authorities. This may be followed by appropriate notifications and/or instructions to the passengers, while the vehicle may also implement respective actions.

While a commuter is in the shuttle a petty crime takes place in the form of assault (Figure 39 – video: https://drive.google.com/open?id=15tLAGcjp5pzI5erHRcg0u2Mh53PSelkt). Also, the commuter is attacked by another person who is attempting to snatch his bag (Figure 40 - video: https://drive.google.com/open?id=1fGktDygbtxWT20ILfqBXPs3fdy8mXozi). The system identifies the event and sends a security alert to the operator or security supervisor. The course of action of the operator is a human decision, meaning, whether he/she will decide to stop the minibus or will notify the passengers via the radio system.

A young person onboards the shuttle during an itinerary performed by the vehicle during the night hours (Figure 41: https://drive.google.com/open?id=125B7YX1w_yKqvr6zyiai22urGXK0nV0j). The youngster attempts to perform a vandalism action on the shuttle, through painting a graffiti on the windows or smashing the window. The event is detected and the person is warned by the radio system of the shuttle or security personnel intervenes by stopping the shuttle.

Inside the shuttle there is unaccompanied luggage which remains as is unmoved for a long time (Figure 42: https://drive.google.com/open?id=1rCV0nYH4LQwOv4hlw1vyeXpXqcdrByMw). In case that the total time of the detected luggage that remains unmoved in the vehicle passes the predefined time frame, a notification is sent to the security operator. The security operator monitors the clips that are captured and evaluates the criticality of the situation and whether to intervene or not.

We compiled a 3 min single video which includes detections from all the above scenarios (video link: https://drive.google.com/drive/folders/1rBm3Ss2XuwbZhiVm5MS8QRYFMuNYrrVu)

Table 5. Precision, Recall and F1-Score metrics for the two classes

Classification Report				
Class	***Precision***	***Recall***	***F1-Score***	***Support***
Normal	0.99	0.99	0.99	1208
Abnormal	0.93	0.95	0.94	147
accuracy	0.99			1355
macro avg	0.96	0.97	0.96	1355
weighted avg	0.99	0.99	0.99	1355

Spatiotemporal Autoencoder results

In order to visualize the predictions, we provide the preprocessed input frame for the current moment at the bottom-left (Figure 44 and 45). The frame is resized from 64 ×64 and grayscale and a mask overlay is obscuring the out-of-interest areas (road). At the right next to the input frame, the resized output of our model is shown. The third mini-frame demonstrates only the significant differences between the input and the output frames with white pixels which are then merged above the original frame for demonstration purposes.

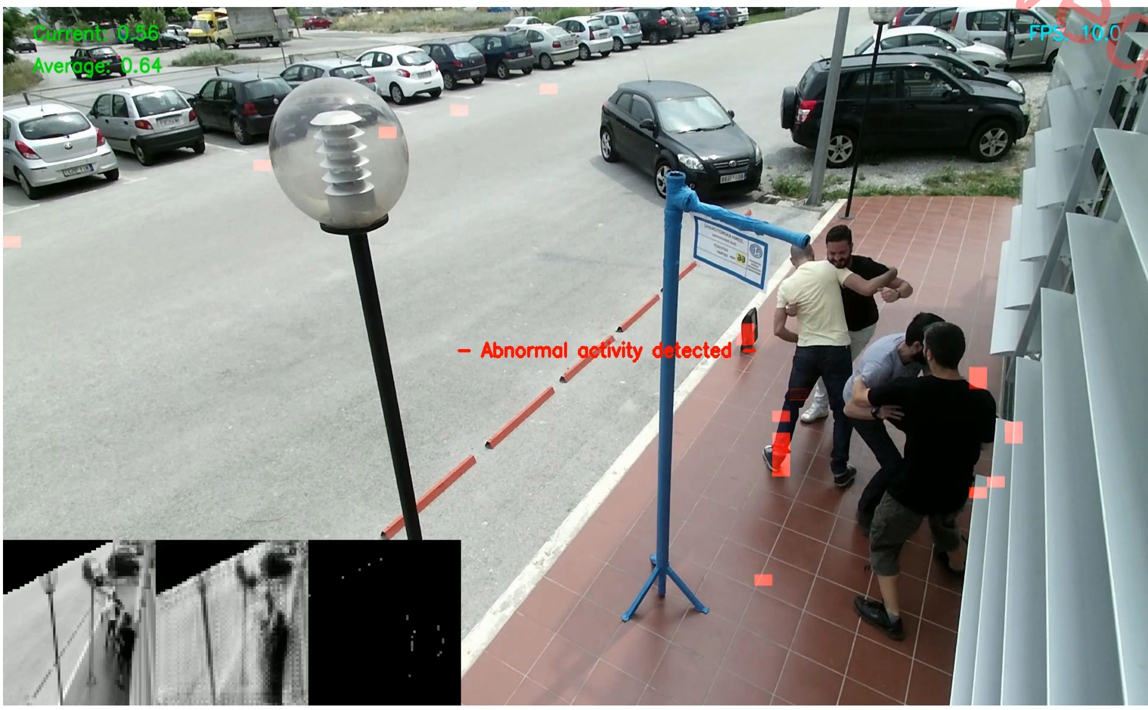


Figure 44. Evaluation on Use Case 2 – Fighting Scenario (1)

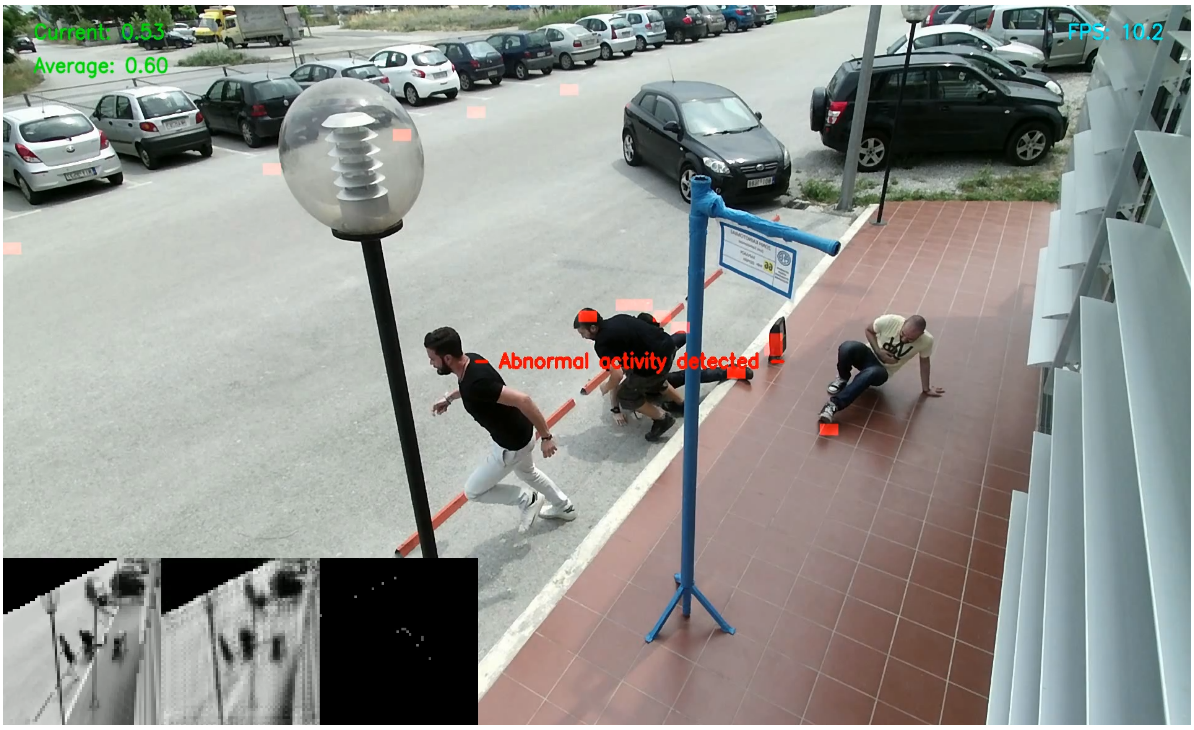


Figure 45. Evaluation on Use Case 2 – Fighting Scenario (2)

SlowFast algorithm results

In Figure 46, we present some results on each use case. Each row contains two images of fighting, bagsnatch, vandalism and falldown events correspondingly.

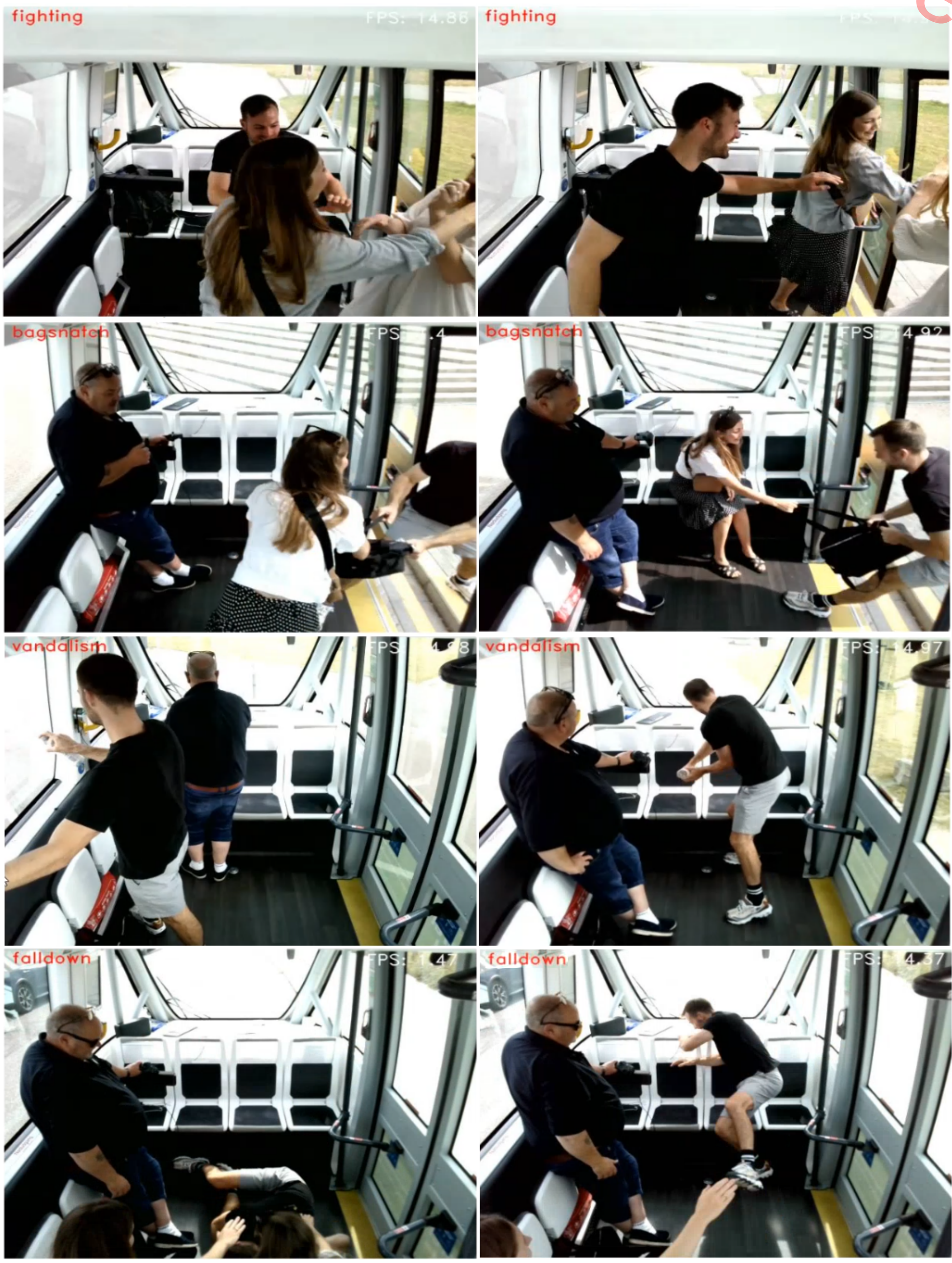


Figure 46. Evaluation on use cases with the SlowFast algorithm. Each row contains two images of fighting, bagsnatch, vandalism and falldown events correspondingly.

Performance/energy analysis

Since we are targeting an autonomous platform, we need to evaluate the efficiency of this system in terms of performance and power requirements. We obtained the approximate GPU power consumption (Figure 46) using the driver tools provided by Nvidia. We estimated a maximum of 100W overhead by the

CPU and other components of the system. In general, it would be safe to assume that our system's energy requirements on full speed peaks at 300 W in total at the hardware. The maximum frame rate achieved on a single Tesla K40 m is about 13 fps (77 ms per frame). Experiments indicated that locking the framerate at a slightly lower value (10 fps) yields much more stable performance and reduces the total power consumption.

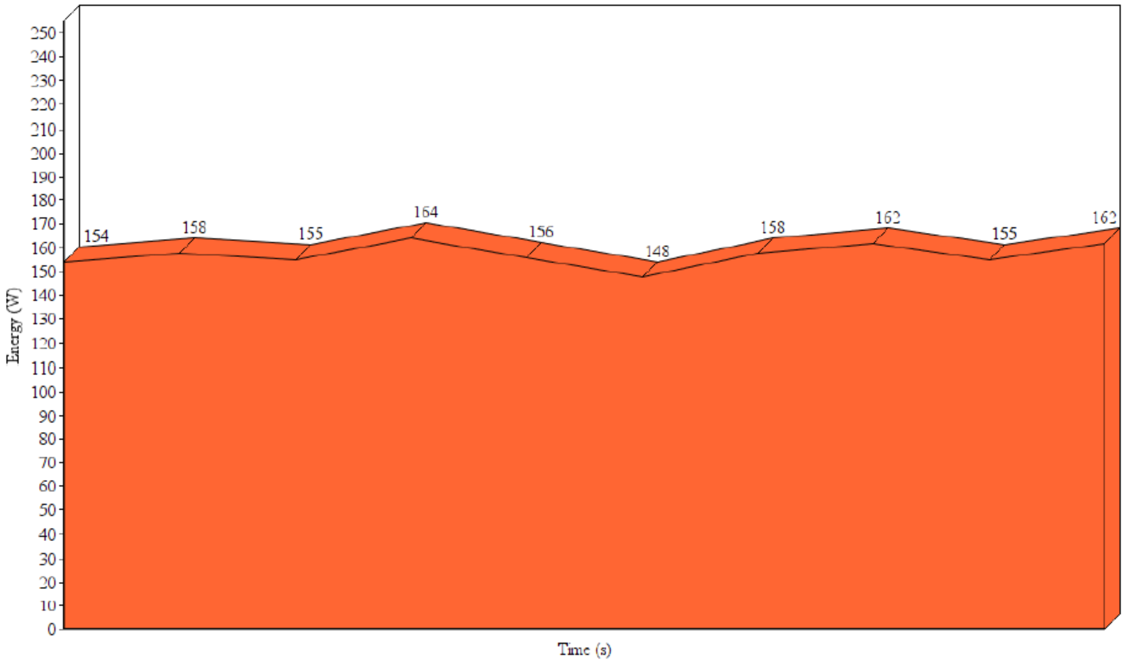


Figure 47. NVIDIA Tesla K40 m GPU energy consumption on full usage (avg. 13 fps)

Sound analysis results

Regarding our experiments, we used a batch size of 16 images and set the initial epochs to 200. However, we noticed that for the DenseNet-121 only eight epochs were sufficient to achieve the optimal performance. We applied an Early Stopping function, where we checked the validation F1 macro averaged score for an improvement in five consecutive epochs. If no improvement was detected the network stopped the training in order to avoid overfitting.

The DenseNet-121 achieved a training F1-Score of 95.92% and a validation F1-Score of 88.74% (Figure 40-left) for the case of 0 dB SNR. Regarding the case of 30 dB, the DenseNet-121 achieved a training F1-Score of 96.84% and a validation F1-Score of 91.82% (Figure 40-right). As expected, the network is able to classify the target classes with higher accuracy in the case of 30 dB SNR and we notice that the training loss starts at smaller values, compared to the case of 0 dB SNR.

There are plenty of options to tweak in a CNN. The number of convolutional layers, max-pooling layers, filter sizes, activation functions. In our experiments, we focus on comparing the default DenseNet-121 architecture in various SNR settings.

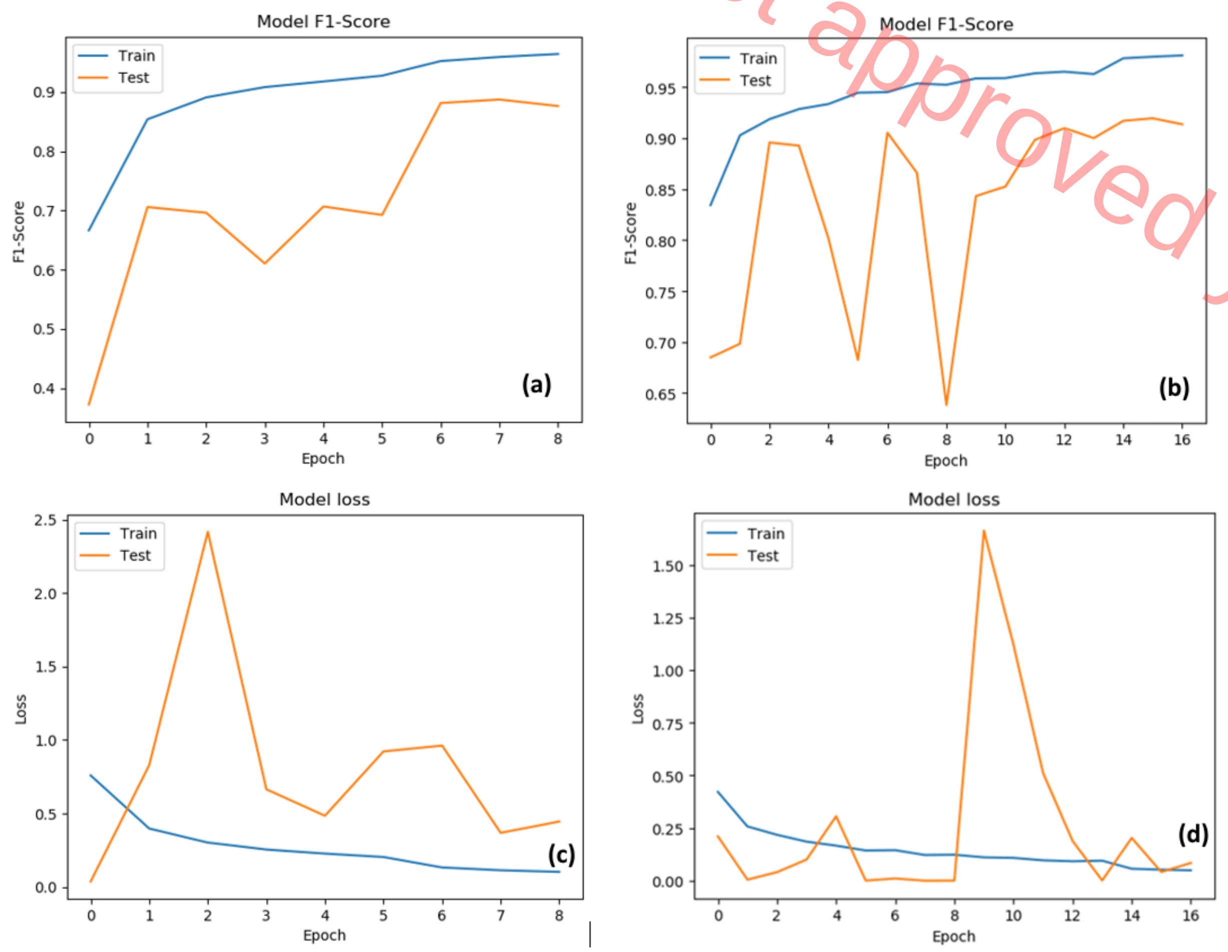


Figure 48. F1-Score (a) and categorical cross-entropy loss (c) for the 0 dB case of the DenseNet-121 architecture. F1-Score (b) and categorical cross-entropy loss (d) for the 30 dB case of the DenseNet-121 architecture

Since the train and validation losses and F1-Scores are not sufficient for the complete evaluation of the proposed framework, we evaluated the Precision, Recall, and F1-Score for each class (Table 6) and calculated the receiver operating characteristic (ROC) curves for each class (Figure 41).

Table 6. Precision, Recall and F1-Score metrics for the four classes at different SNR levels

Classification Report		Classes											
		Background Noise			**Glass Breaking**			**Gun Shot**			**Scream**		
		Precision	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
SN	**-5 dB**	79 %	86 %	83 %	74 %	82 %	78 %	89 %	87 %	88 %	85 %	60 %	70 %

R values	**0 dB**	87 %	90 %	89 %	86 %	90 %	88 %	97 %	96 %	96 %	88 %	77 %	82 %
	5 dB	91 %	85 %	88 %	80 %	94 %	87 %	92 %	94 %	93 %	88 %	87 %	87 %
	10 dB	89 %	91 %	90 %	89 %	90 %	90 %	99 %	97 %	98 %	88 %	83 %	86 %
	15 dB	93 %	89 %	91 %	88 %	93 %	90 %	97 %	98 %	98 %	88 %	90 %	89 %
	20 dB	92 %	90 %	91 %	88 %	93 %	90 %	99 %	99 %	99 %	89 %	89 %	89 %
	25 dB	91 %	91 %	91 %	87 %	89 %	88 %	99 %	99 %	99 %	91 %	89 %	90 %
	30 dB	92 %	90 %	91 %	88 %	90 %	89 %	99 %	99 %	99 %	88 %	91 %	90 %

From Table 6 we can see that while the SNR value increases, the network is able to distinguish the four classes more accurately. Specifically, for all four classes, the network achieves the highest F1-Score for SNR values higher than 15 dB.
The class “gun shot” is one of the most easily distinguishable, while the “scream” class is the hardest to classify.

Regarding the multichannel spectrogram representations, a representative sample of the variation with respect to each class and SNR in the extended dataset is shown in Figure 47. It is evident that, as the SNR increases, the features become clearer and easier to distinguish from the background, as was expected. Although the focus of the present study is mainly on multichannel spectrogram representation performance (as well as studying different single channel representations) and the study of the effect of the SNR on the performance, a comparison of different studies conducted on the MIVIA Audio Events dataset is shown in Table 7. The two common representations mentioned in the literature are spectrograms and gammatonegrams. The former is the traditional time-frequency visualisation, but it actually has some important differences from how sound is analysed by the ear; most significantly, the ear’s frequency sub-bands get wider for higher frequencies, whereas the spectrogram has a constant bandwidth across all frequency channels. A Gammatone spectrogram or gammatonegram is a time-frequency magnitude array based on an FFT-based approximation to gammatone sub-band filters, for which the bandwidth increases with increasing central frequency. The upper part of the table compares the results achieved by considering the classification of positive SNR sound events only are shown. In the lower part of the Table, the results achieved by including sound events with negative and null SNR to the above are exhibited. The average RRs for the three classes of interest (event-based) were 92.5% and 90.9% for the original and the extended dataset, respectively. The latter compares well with the reported value of 90.7% in[38].

[38] Strisciuglio, N.; Vento, M.; Petkov, N. Learning representations of sound using trainable COPE feature extractors. Pattern Recognit. 2019, 92, 25–36.

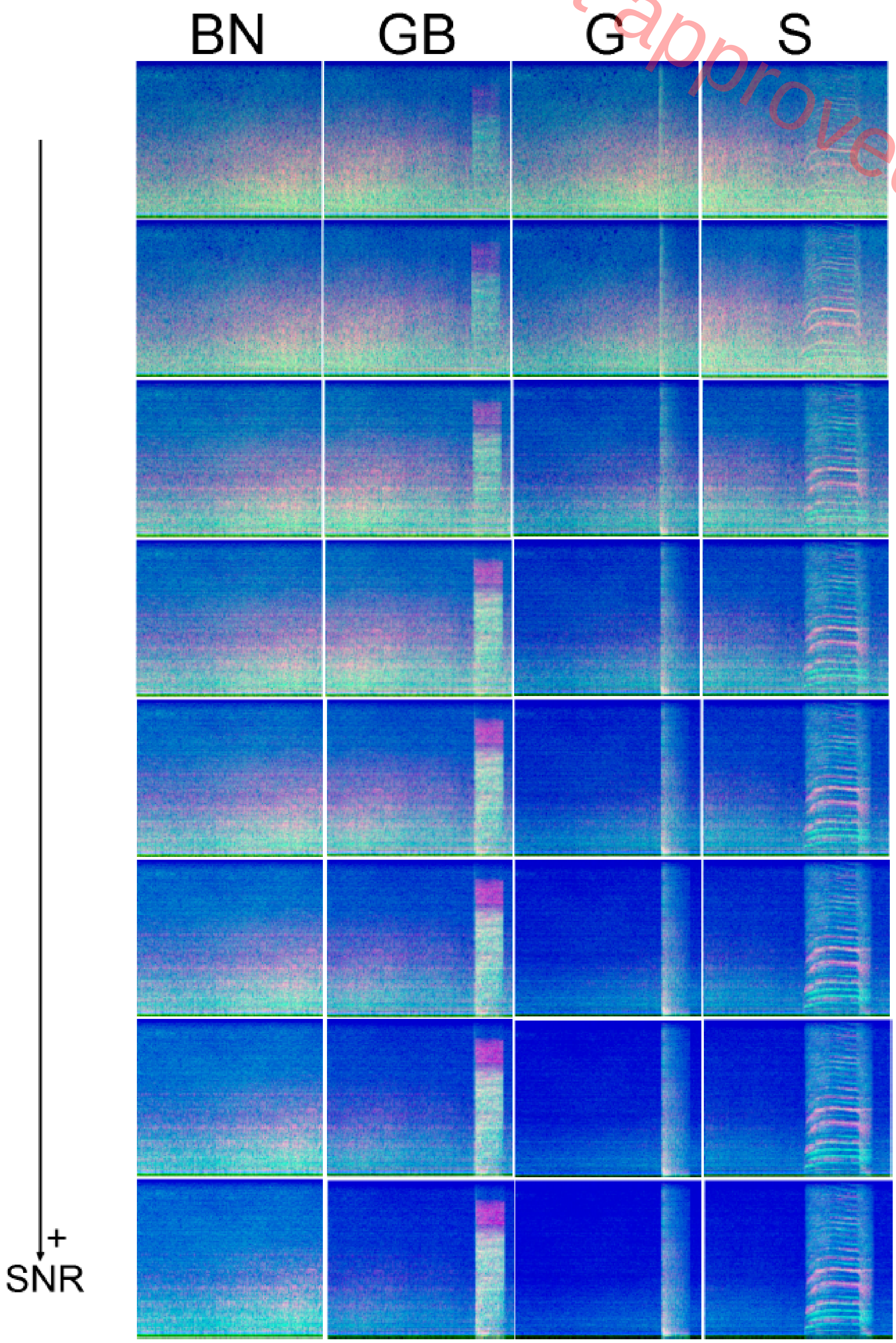


Figure 49.Multichannel spectrograms obtained with the stacked method discussed in Section 3.3.1.2.: the representations of the four classes of the extended dataset are shown, with the SNR value increasing at the direction of the arrow (left side) from -5 dB to 30 dB with a step of 5 dB.

Table 7. Results (frame-by-frame) of available studies in the literature along with the results of the current work, regarding the four classes (including the background noise) of the original and the extended MIVIA Audio Events Dataset.

Method Test with SNR > 0	Representation	RR (%)	Accuracy (%)	MDR (%)	ER (%)	FPR (%)
Present Study	STFT + Mel + MFCC (Stacked)	92.5	95.21	7.28	0.22	2.59
SoundNet[38]	Gammatonegram	93.33	-	9.9	1.4	1.4

Finally, in order to test the generalizability of the selected DenseNet architecture, we tested the network under three settings. The first one with the network trained on 30 dB SNR and tested on 0 dB SNR. The second one with the network trained on 0 dB SNR and tested on 30 dB SNR and finally with the network trained on 15 dB SNR and tested on 30 dB SNR. The classification reports are summarized in Figures 50 – 52.

```
Classification Report
                  precision    recall  f1-score   support

Background Noise       0.48      0.99      0.64      2358
  Glass Breaking       0.80      0.11      0.20       900
        Gun Shot       1.00      0.00      0.00       900
          Scream       0.82      0.04      0.07       900

        accuracy                           0.49      5058
       macro avg       0.78      0.29      0.23      5058
    weighted avg       0.69      0.49      0.35      5058
```

Figure 50. Classification report for DenseNet-121 trained on 30 dB and tested on 0 dB SNR

```
Classification Report
                  precision    recall  f1-score   support

Background Noise       0.71      0.84      0.77      2358
  Glass Breaking       0.80      0.96      0.88       900
        Gun Shot       0.97      0.28      0.44       900
          Scream       0.83      0.86      0.84       900

        accuracy                           0.77      5058
       macro avg       0.83      0.74      0.73      5058
    weighted avg       0.79      0.77      0.74      5058
```

Figure 51. Classification report for DenseNet-121 trained on 0 dB and tested on 30 dB SNR

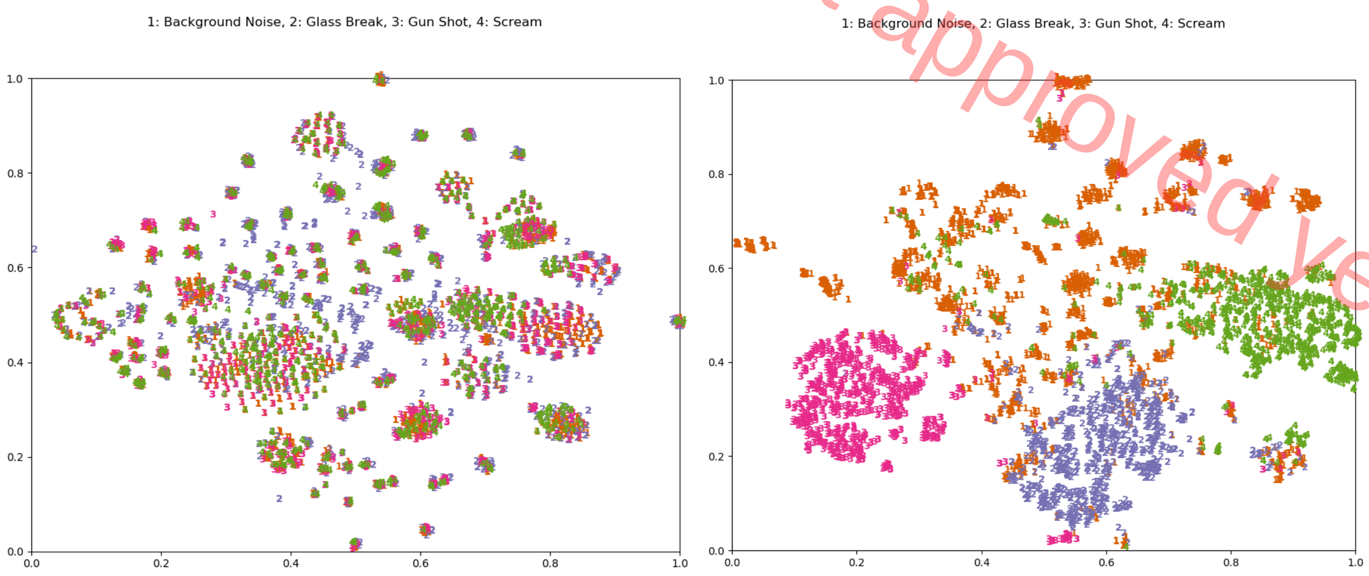
Classification Report

	precision	recall	f1-score	support
Background Noise	0.95	0.84	0.89	2358
Glass Breaking	0.82	0.98	0.89	900
Gun Shot	0.98	0.99	0.99	900
Scream	0.84	0.92	0.88	900
accuracy			0.91	5058
macro avg	0.90	0.93	0.91	5058
weighted avg	0.91	0.91	0.91	5058

Figure 52. Classification report for DenseNet-121 trained on 15 dB and tested on 30 dB

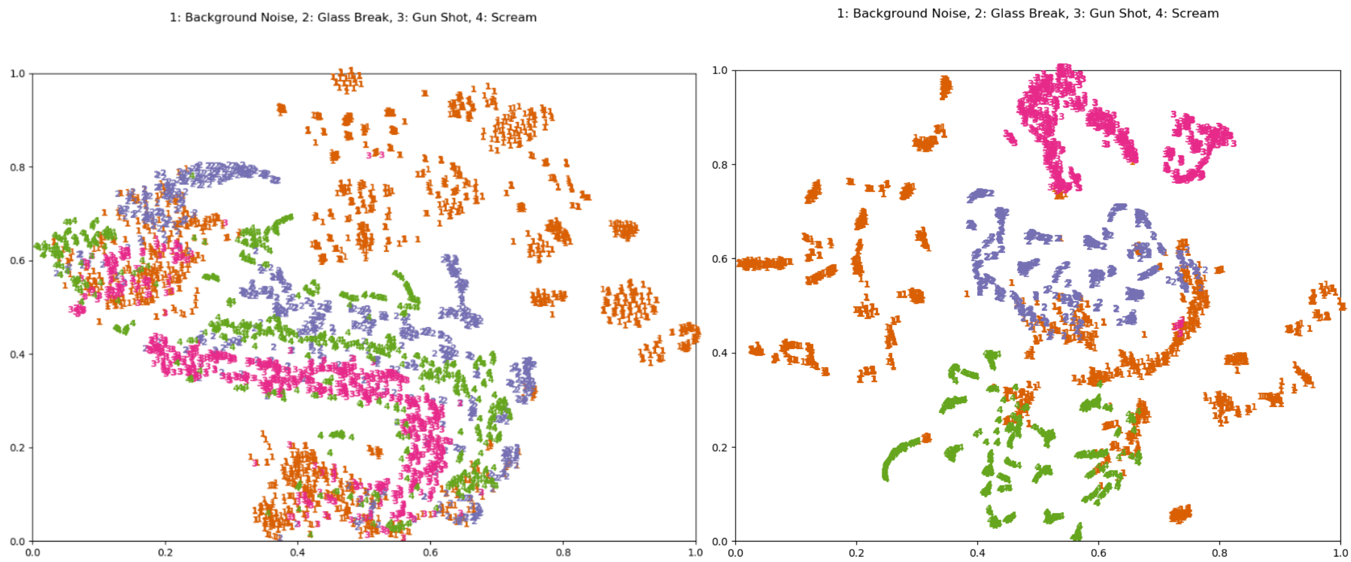
From Figure 48, we notice that the network trained in an environment with the least noise cannot distinguish the classes in a noisy environment. On the other hand, the network trained on a noisier environment (Figure 50) can distinguish the classes in the quietest environment settings almost as well as the network trained on the same environmental settings. Therefore, we notice that our network can generalize well in clean environments when trained in noisy ones.

Despite the classification reports and the ROC curves, we used the T-distributed Stochastic Neighbor Embedding (t-SNE) plots to further visualize the automatic features that were learnt by the proposed 2D CNN architecture. The advantage of the t-SNE visualization, compared to a Principal Component Analysis, is that it uses the local relationships between points to create a low-dimensional mapping. This allows the t-SNE to capture non-linear structure of the given dataset (raw data and learnt features), since the neural network is learning non-linear representations of the dataset. Figure 45 shows the t-SNE visualization at 0 dB (noisy environment) and Figure 46 shows the t-SNE visualization at 30 dB (quiet environment).



(a) (b)

Figure 53. t-SNE results at 0 dB



(a) (b)

Figure 54. t-SNE results at 30 dB

From the above Figures we notice the randomness in the 2D space of the raw features, in both environments and on the right part of the Figures the ability of the proposed 2D CNN architecture to distinguish between the target classes and create clear clusters for each class.

Since there is not a public dataset with similar sound events in a shuttle environment for in-cabin monitoring services, the proposed architecture had to be evaluated on data collected in the HOLO test track. A small dataset, as an initial phase, consisting of 200 3-second clips with ambient noise and 75 3-second events with female and male screams was recorded. All 275 sound clips were successfully classified and were not confused with glass breaking or gun shots. Collecting real data from the latter classes was not possible in this project. Additional details, with unseen during the training datasets, can be found in the paper by Papadimitriou, Vafeiadis, et al[39].

Microphone distance, ambient noise and SNR are well-known challenges in audio analysis and classification; they are factors that differentiate the latter from fields that have proven more straightforward for image analysis. The present analysis aim was to tackle the aforementioned issues and to provide a form of analysis that generalizes well even when background noise is high and/or the signal of the event of interest is weak and the SNR drops to the negative territory.

Before deploying the system for the real-life application, more experiments will be conducted with other network architectures (e.g., other 2D CNNs, 1D CNNs, RNNs), ensemble methods will be explored and statistical test will be performed in order to determine whether the results are independent or there is a degree of randomness.

3.3.2 Equipment

At previous Sections, we presented three different solutions for the vision modality. In this section, we are going to compare those solutions in terms of requirements and equipment and enlist the advantages and drawbacks of each method.

Stacked LSTM classification is a supervised method that classifies the extracted skeleton key points based on a (supervised) dataset. It can reliably detect various types of events but depends on a previous skeleton extraction and tracking process, which may not be accurately feasible due to space and occlusion constraints. It does not depend on the camera setup (but it is less effective on top-down cameras with a wide field of view). The Convolutional LSTM autoencoder (unsupervised) extracts spatiotemporal features from a video sequence, able to learn regularity. It can detect abnormal events and activities and depends on the camera setup but due to its unsupervised nature it can be trained over-time and self-improve. It is ideal for crowded areas and static camera setups. In Hybrid classification (semi-supervised), the previous encoder acts as a high recaller, and the anomalies are sent through a false positive reduction model (hybrid model). This combination provides a deep neural network with high recall and high precision.

We developed different solutions due to the perspective of the final camera setup and the technical specifications of the lens/sensor (Table 8). Existing in-shuttle camera have a wide-angle field of view and a top-down perspective which is not suitable for pose estimation. In some occasions the whole body of the commuter can be occluded by his head, rendering impossible for the pose estimation algorithms to detect his pose in 2D. Although the LSTM Classification via Pose Estimation performs generally better, it cannot be applied with satisfactory results in the existing in-shuttle camera. We developed the autoencoder based solutions to prevent excluding this possibility. Autoencoders do not depend on pose estimation and extract their own features by learning regularity.

Taking the above into account, the relevant equipment for each case can be derived. The choice of the final solution is based on these characteristics, as well as on the availability of connection with the system and real-time data frames.

[39] Papadimitriou, I.; Vafeiadis, A.; Lalas, A.; Votis, K.; Tzovaras, D. Audio-Based Event Detection at Different SNR Settings Using Two-Dimensional Spectrogram Magnitude Representations. Electronics 2020, 9, 1593. https://doi.org/10.3390/electronics9101593

Table 8. Equipment and proposed solutions

Requirements	Custom installation by CERTH	Integration to NAVYA's equipment
Solution	LSTM Classification via Pose Estimation	Spatiotemporal Autoencoder/Hybrid LSTM Classification
Camera	2x FullHD camera sensors (require installation)	Existing in-shuttle camera
Camera Power supply	Using USB protocol (requires wiring)	Already exists
Data transmission	Using USB protocol (requires wiring)	Requires access to the video stream via a standard protocol (e.g RTSP) or a documented API
Host PC	In-shuttle PC capable of real-time inference	In-shuttle PC capable of real-time inference
PC Power supply	1000W (max)	1000W (max)

3.4 Prototyping plan

The prototyping plan of this service consists of the development phase and the deployment phase:

Development phase: In this phase, CERTH initially plans and analyses the requirements with inputs from all the stakeholders. Once the requirement analysis is done, the final software and hardware representation and documentation of the requirements are accepted from the project stakeholders. The next stage consists of designing the critical service components, the research and the development of the algorithms and the integration of the components across several platforms. Also, in this stage several data captures are performed for training the machine learning algorithms. CERTH will perform additional data capture sessions with different lighting conditions and operators. The captured data should demonstrate multiple scenarios, both with abnormal and normal events, and various capturing conditions. Also, this phase includes the re-evaluation of the results, using the new data acquired for retraining the machine learning algorithms.

Deployment phase: In the deployment phase, the individual service components are unified into a complete system. The system has discrete input and output and is ready to be integrated with other platforms. In the following months, CERTH will deploy the service in a demo AV, perform internal tests and verifications with all the stakeholders. In this stage, minor modifications and fine-tuning may be applied on the final setup, mainly on design or algorithms of the service, depending on the real operation conditions.

Prototyping phase: Amobility will in coordination with CERTH install cameras and sensors in the Amobility shuttles for testing and validation of the technologies, use cases etc.

3.5 Result analysis

In the evaluation phase, the service will be tested under real conditions that can be performed on controlled routes of Amobility along with a safety officer inside the shuttle, depending on the GDPR permissions. Also, short-sessions for assessment with real passengers will be available. After the successful evaluation of the service in Amobility sites, the results will be used to fine tune the service if required and the service will be deployed and evaluated also in other sites of the involved operators in the AVENUE project towards a successful integration to the relevant pilot sites.

4 Service: Automated passenger presence

4.1 Concept of service

The service "Automated passenger presence" aims to address a basic problem of operators' services which is related to the occupation of their vehicles as well as the awareness of the number of people on-board in order to schedule the routes. Furthermore, the passengers would like to know in advance if there is an available seat or enough space on a shuttle to plan their boarding. Traditionally, but also nowadays, passenger counting is conducted manually via passenger surveys or human ride checkers. Typically, the driver or inspectors are responsible for performing enumeration of the onboard passengers, something not feasible in an autonomous shuttle. Automatic passenger counting has been rapidly emerging in recent years to address similar needs. An automated system is introduced capable to detect passenger presence in real-time with high accuracy, count onboard passengers and calculate vehicle occupancy. Surveillance using sensors such as cameras (cameras of different technologies can be used so that passengers' privacy is protected) and smart software in the bus will automate the detection of passenger presence.

Several concerns of the end users regarding the Safety and Robustness of the autonomous vehicles that are directly linked to the final User Acceptance of the new technology, can be identified. The prospective passengers may deal with several possible instances that could arise in case there is no staff inside the shuttle. Indicatively:

- No one will be in the bus to count the number of passengers with regard to the shuttle's capacity
- Continuous stops throughout the entire route, even in cases where the shuttle is fully occupied
- No authority figure present to alert passengers of their designated bus stop

To address the aforementioned concerns on social and personal safety and quality-of-service into the vehicle, certain measures need to be implemented. For example, counting the number of passengers inside the autonomous vehicle could help avoid overcrowding in the shuttle, as well as meaningless stops in cases where the bus is at full capacity. This may be followed by appropriate notifications and/or instructions to the passengers, while the vehicle may also implement respective actions.

The service provides a video analysis of the vehicle internals, using the on-board camera, in order to identify the vehicle occupation, vehicle free space, as well as counting people on-board. Automatic assessment of space occupation using the on-board cameras is enabled. Capacity is set as an absolute number of space units. For example, each space unit is associated with one standing passenger. Occupancy is set as an absolute number of space units currently in the shuttle. For the operation manager, occupancy is visible on the dashboard of the AVENUE platform, whereas occupancy is displayed as real time information via the AVENUE mobile app, wherever the traveller is. Each passenger (normal, big size, wheelchair user, seated) can determine whether he/ she can fit in or not. Assessment for different cases can be provided to assist the passengers on determining whether to request onboarding or not. Automatic counting of people using the on-board cameras is also provided. Additional information can be derived from automated people-counting, while fusion of data related to space occupation and counted number of people can provide more accurate information about the capacity and occupancy of the vehicle. Moreover, occupancy marked with information for the different user cases is displayed as real time information, wherever the traveller is, however it does not guarantee them a free spot by the time the shuttle reaches the station of their choice.

In this section, the implementation of a video analytics software module for an automated passenger presence counting subsystem or for cloud-based services of the system are described along with appropriate planning for the deployment and test of the service into the pilot sites of the AVENUE project.

4.1.1 Use case

The service addresses the timely, accurate, robust and automatic counting of the passenger number within the autonomous vehicle, as well as appropriate notifications and/or instructions to its passengers. The main goal is to monitor the number of passengers onboard at all times, so that the shuttle's capacity is not exceeded at any time. The passenger would like to know if there is an available seat or space before getting on-board on the autonomous shuttle. Occupancy is displayed as real time information, wherever the traveller is, but it does not guarantee them a free spot when the shuttle arrives at the station of their choice. Within the suggested software module, analytics algorithms were developed for the timely, accurate, robust and automatic detection of onboard passengers. The service is able to estimate occupied seats and report the passenger capacity and vehicle occupancy of the autonomous shuttle continuously.

In the context of AVENUE project, the following use cases have been identified to be further examined and addressed:

Use Case 1: Passenger Counting

- The autonomous shuttle has a fixed capacity regarding the number of passengers it can carry.
- The video cameras installed in the autonomous shuttle acquire the color depth images and the data are fed into the system's video analytics algorithms for further analysis.
- In case the algorithms identify that the total number of passengers is reached, the shuttle stops receiving any others and appropriate notifications are sent to the AVENUE mobile app for the passengers that would like to board.

Use Case 2: Route Optimization

- Even though the shuttle is at full capacity, there may still be people waiting ata bus stop to go aboard.
- The bus only makes a stop when a passenger needs to get-off, while the route is modified to save time and cost.
- The number of onboard passengers is always being monitored, so that new passengers could get on, in case of availability.

Use Case 3: Passenger Awareness

- Even though the autonomous vehicle has reached its terminal, there could still be passengers onboard.
- The shuttle counts the number of passengers to make sure there is no one left.
- If there are passengers, the bus alerts them to get-off.

4.2 Stakeholders (development/prototyping team)

For the "Automated passenger presence" service the stakeholders are the same as in Stakeholders (development/prototyping team) (Section 2.4.2).

4.3 Technical requirements

For the "Automated passenger presence" service the technical requirements are the same as in Technical requirements (Service: Security trust services (C)), without the need for any audio sensing equipment (such as microphones).

4.3.1 Technology

Automatic passenger counting systems have evolved considerably within the past 40 years. Passenger flow data can be acquired with high accuracy outperforming manual ride checkers[40]. Devices that operate on 3D image streams are the industrial state-of-the-art technology. Latest generation devices offer an accuracy of around 99% [41] and technical progress is ongoing.

Background

So far, a wide range of competing automatic passenger counting technologies has been developed. Detection methods include infrared light beam cells, passive infrared detectors, infrared cameras, stereoscopic video cameras, laser scanners, ultrasonic detectors, microwave radars, piezoelectric mats, switching mats, and also electronic weighing equipment (EWE)[42].

Operators usually mount one or multiple sensors to collect data in each door area of public transport vehicles like buses, trams, and trains. The number of boarding and alighting passengers are counted separately by converting 3D video streams (infrared beam break) or light barrier methods, which are the most commonly used technologies.

In recent years also weight-based EWE approaches utilizing pressure measurements in the vehicle braking/air bag suspension system have emerged to estimate passenger numbers[43]. These relatively new approaches have proven to provide easy-to-acquire additional information since modern buses and powertrains are equipped with (intelligent) pressure sensors by default. However, those methods are considered not feasible in smaller scale vehicles, such as the autonomous buses in our case, due to design constraints and other restrictions.

Passenger Detection

As depicted in Figure 55, the first layer of sensors connects to the Hardware Abstraction Layer (HAL). The HAL implements the IP and the USB protocol supporting IP and USB cameras correspondingly. The input data are converted and transformed in a compatible format and passed into the analytics algorithms. The prediction is then transferred via the API endpoints into the cloud. Other systems and services have access to data.

[40] Hwang, M., Kemp, J., Lerner-Lam, E., Neuerburg, N., Okunieff, P.E.: Advanced public transportation systems: the state of the art update 2006. FTA report (2006)

[41] Hella Aglaia: Public Transport: HELLA Aglaia People Sensing (2018). http://people-sensing.com/public-transport.

[42] Kotz, A.J., Kittelson, D.B., Northrop, W.F.: Novel vehicle mass-based automated passenger counter for transit applications. Transp. Res. Record 2536, 37–43 (2015)

[43] Nielsen, B.F., Frølich, L., Nielsen, O.A., Filges, D.: Estimating passenger numbers in trains using existing weighing capabilities. Transportmetrica A Transp. Sci. 10(6), 502–517 (2014)

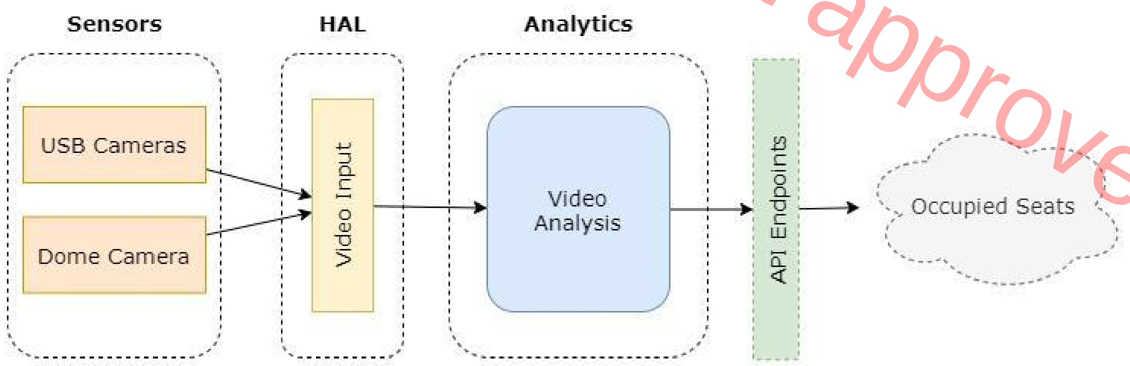


Figure 55. High level overview of the "Automated passenger presence" service

In order to detect timely and accurately human items, we implemented YOLO[44], a state-of-the-art and real-time object detection system. YOLO applies a single neural network to the full image, which divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities.

YOLO model has several advantages over classifier-based systems. It looks at the whole image at test time so its predictions are informed by global context in the image. It also makes predictions with a single network evaluation unlike systems like R-CNN which require thousands for a single image. This makes it extremely fast, more than 1000x (Figure 57) faster than R-CNN and 100x faster than Fast R-CNN.

Methodology

Dataset: We collected and labelled a new dataset named CERTH - AVenue Overhead Fisheye (C-AVOF). The new dataset contains frames and human objects in a simulated shuttle environment, and also includes challenging scenarios such as crowded rooms, various body poses, and various-light conditions. The camera used for the data capture process is the D-Link DCS-4625 at 1080p resolution output. During the annotation of this dataset, we used Deepsort[45] for tracking the individual passengers in the shuttle across multiple frames and thus our dataset can be also used for additional vision tasks using overhead, fisheye images, such as video-object tracking and human re-identification. Also, for evaluation purposes, some cherry-picked samples from the BOSS dataset[46] were included, especially the scenarios from camera 5 and 7 with the top-down overhead perspective.

Network architecture: Inspired by RAPiD[47], the passenger detection network consists of three stages: the backbone network, the feature pyramid network (FPN), and the bounding box regression network. The backbone network works as a feature extractor that takes an image as input and outputs a list of features

[44] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[45] Wojke, N., Bewley, A.: Deep cosine metric learning for person re-identification. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 748–756. IEEE (2018)

[46] Velastin, S.A., G´omez-Lira, D.A.: People detection and pose classification inside a moving train using computer vision. In: Zaman, H.B., et al. (eds.) International Visual Informatics Conference, pp. 319–330. Springer, Cham (2017). https://doi. org/10.1007/978-3-319-70010-6 30

[47] Duan, Z., Tezcan, O., Nakamura, H., Ishwar, P., Konrad, J.: Rapid: rotation-aware people detection in overhead fisheye images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 636–637 (2020)

from different parts of the network. In the next stage, we pass those features into the FPN, in order to extract features related to object detection. Finally, at the last stage, a Convolutional Neural Network (CNN) is applied to each feature vector in order to produce a transformed version of the bounding-box predictions (Figure 56).

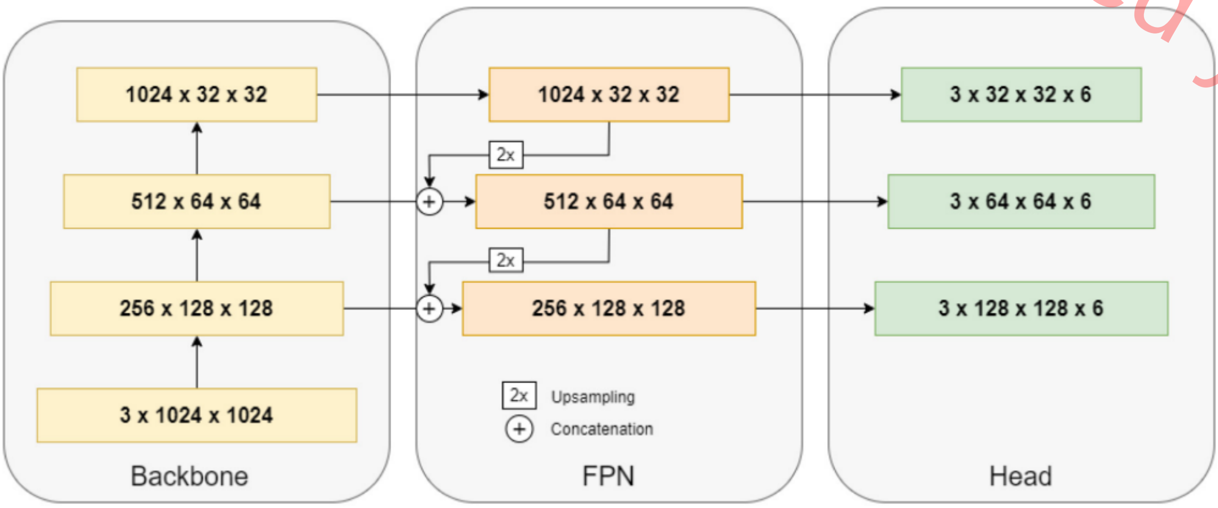


Figure 56. An illustration of multiple convolutional layers and multi-dimensional matrices such as the feature maps with 1024×1024 input resolution

Training: We used the pretrained RAPiD weights on MW-R and HABBOF which we fine tuned by training on our custom dataset C-AVOF. Rotation, flipping, resizing, and color augmentation are used in the training stage.

Post-processing: After the detection process, we used the bounding box coordinates in order to compute the centroid. In order to calculate the distance between each bounding box, we used the Euclidean distance formula. Each bounding box is connected with the rest via distance line which represents the real-world distance, when multiplied with a weight factor. The weight value is calculated via the camera calibration process and takes into account various parameters such as the position of the camera and the field of view.

4.3.1.1 Research results

The implemented algorithm monitors the shuttle in real time and records the position of each person in frames (Figure 58).

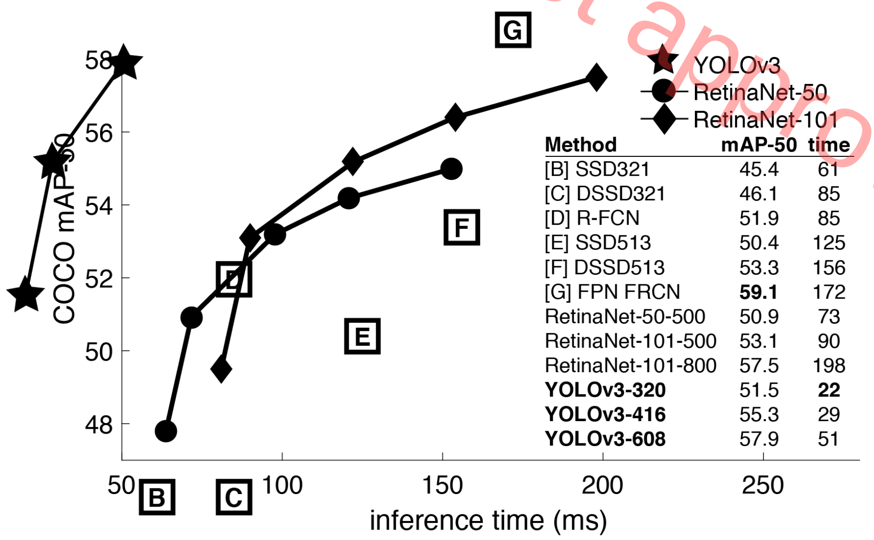


Figure 57. Performance comparison between various state-of-the art methods[48]

Figure 58. People counting using the on-board camera by NAVYA

More results are also presented in the Video analysis results sub-section of the “Enhance the sense of security and trust” service.

The experiments were performed on the Nvidia Jetson AGX Xavier platform that features 512-core Volta GPU with Tensor Cores, 8-core ARM 64bit CPU and 32 GB 256-Bit LPDDR4x RAM. The camera system is the D-Link DCS4625 panoramic fisheye camera with f/2.0 wide-angle panoramic lens and 5MP (2560 × 1920 @30 fps) high-resolution video. Results in Table 9 indicate that RAPiD at 608 × 608 resolution achieved the best performance and the fastest execution speed on our dataset C-AVOF. The averaged FPS value represents the execution speed on the Nvidia Jetson AGX Xavier.

[48] https://pjreddie.com/darknet/yolo/

Table 9. Performance comparison of two state-of-the-art methods on our dataset C-AVOF

	FPS	*MW-R*[49]	*HABBOF*[50]	*CEPDOF49*	*C-AVOF (ours)*
		AP_{50}	AP_{50}	AP_{50}	AP_{50}
Tamura et. al. [51] (608)	5.8	77.4	86.1	59.2	77.6
RAPiD (608)	6.5	96.2	97.6	84.3	97.9

In Figure 59 and Figure 60, we can see the visualized results on unseen scenarios from the BOSS dataset. The blue bounding boxes indicate the detections, and the numbers represent the confidence of each detection. Red lines denote an unsafe distance while lines in green a sufficient distance.

[49] Duan, Z., Tezcan, O., Nakamura, H., Ishwar, P., Konrad, J.: Rapid: rotation-aware people detection in overhead fisheye images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 636–637 (2020)

[50] Li, S., Tezcan, M.O., Ishwar, P., Konrad, J.: Supervised people counting using an overhead fisheye camera. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8. IEEE (2019)

[51] Tamura, M., Horiguchi, S., Murakami, T.: Omnidirectional pedestrian detection by rotation invariant training. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1989–1998. IEEE (2019)

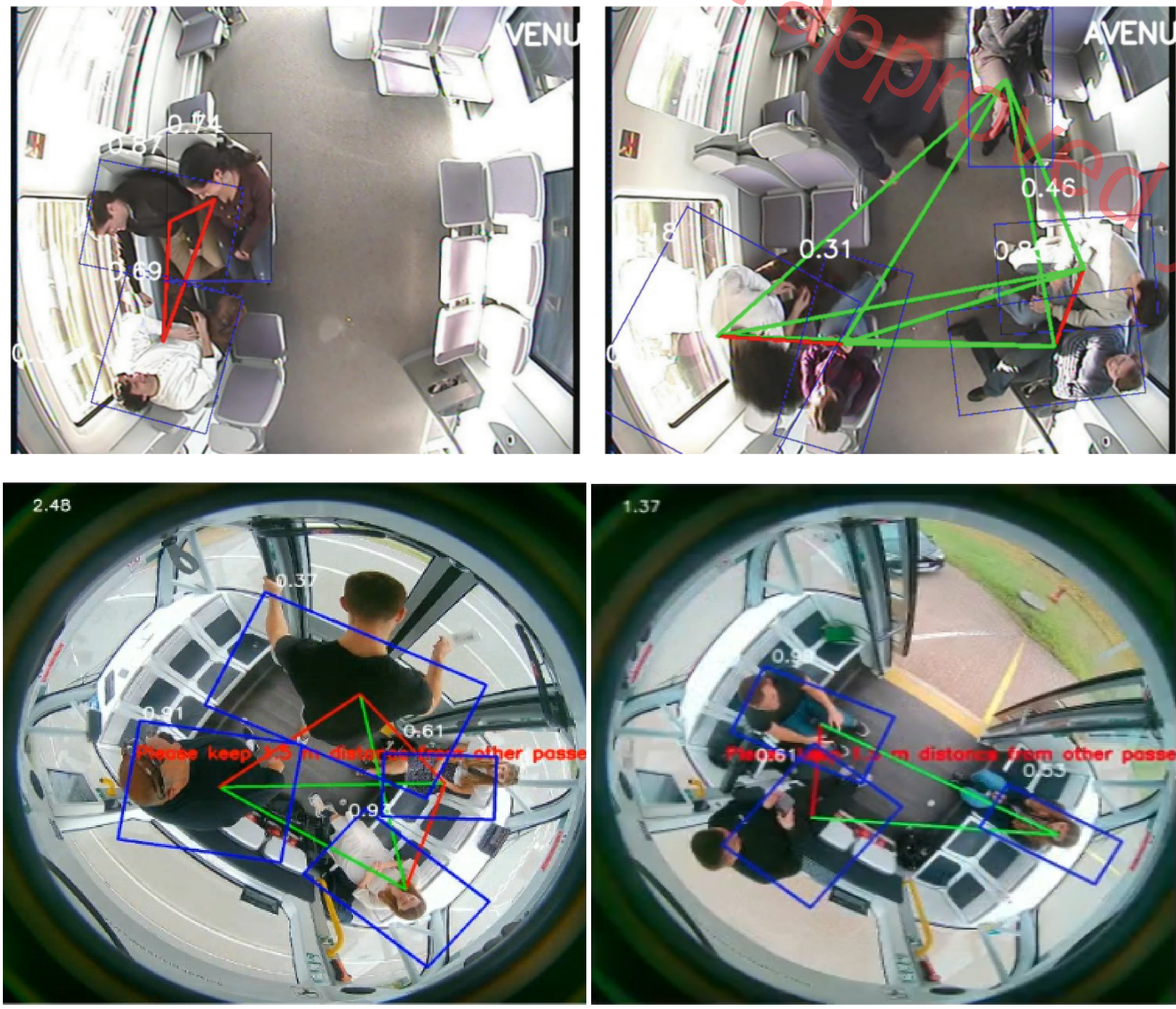


Figure 59. Results on unseen scenarios from the BOSS and HOLO+TPG datasets. Green lines represent a safe distance while red lines an unsafe one.

Figure 60. Results on an unseen crowded scenario from the BOSS dataset, with overlapping detections.

Handling reflections is still a challenge in certain scenarios, as also happens in similar approaches. Passenger figures might appear to the windows of the shuttle as reflections, especially when the lighting is low. To mitigate this issue, a custom mask is applied on large reflective surfaces.

Adaptive inference: As the service can report real-time metrics over the network to the AVENUE dashboard, we exploit this bidirectional communication inside the shuttle ecosystem, to enhance its awareness. By receiving useful data about the vehicle's state, such as the speed, the status of the door and the tires pressure, we can reduce the power consumption and be more energy efficient by decreasing the number of inferences, as Table 10 indicates.

Table 10. Estimated power consumption between adaptive modes measured on Jetson AGX Xavier using tegrastats utility

Inference frequency	*Power consumption (Wh)*
Continuous (real-time)	26.2
Adaptive (crowded-multiple stops)	14.7
Adaptive (average)	11.4

4.3.2 Equipment

The same equipment is utilised that is enlisted in Section 2.4.3.2 and 2.4.3 of the Security trust service without the need for any audio sensing equipment (such as microphones).

4.4 Prototyping plan

The prototyping plan of this service consists of the development phase and the deployment phase:

Development phase: In this phase, CERTH initially plans and analyses the requirements with inputs from all the stakeholders. Once the requirement analysis is done, the final software and hardware representation and documentation of the requirements are accepted from the project stakeholders. The next stage consists of designing the critical service components, the research and the development of the algorithms and the integration of the components across several platforms. Also, in this stage several data captures are performed for training the employed machine learning algorithms. CERTH will perform additional data capture sessions with different lighting conditions, angles and operators. The same data from the "Enhance the sense of security and trust" service's Prototyping plan can be reused. An initial setup of the equipment and the camera sensors is required, along with the interconnection of the components and software preparation. The captured data should demonstrate multiple camera angles and various capturing conditions. The next steps include the re-evaluation of the results, using the new data acquired for retraining the machine learning algorithms.

Deployment phase: In the deployment phase, the individual service components are unified into a complete system. The system has discrete input and output and is ready to be integrated with other platforms. In the following months, CERTH will deploy the service in a demo AV, perform internal tests and verifications with all the stakeholders. In this stage, minor modifications and fine-tuning may be applied on the final setup, mainly on design or algorithms of the service depending on the real operation conditions.

Prototyping phase: Amobility will in coordination with CERTH install cameras and sensors in the Amobility shuttles for testing and validation of the technologies, use cases etc.

4.5 Result analysis

At the final step, an evaluation of the service under real conditions follows, that can be performed on daily or controlled routes of Amobility operator along with a safety officer inside the shuttle, depending on the GDPR permissions. Also, short-sessions for assessment with real passengers will be available. The results will also be compared with the manual counting done by Amoblity for validation purposes of the service's accuracy. After the successful evaluation of the service in Amobility sites, the results will be used to fine tune the service if required and the service will be deployed and evaluated also in other sites of the involved operators in the AVENUE project towards a successful integration to the relevant pilot sites.

5 Service: Follow my kid/grandparents

5.1 Concept of service

The service "Follow my kid/grandparents" is designed to increase autonomy of non-fully autonomous people (Kids, Grandparent(s), disabled people etc.). It will allow carers or family members to be sure that their beloved family members are safe while moving around the city using public transports. On the other hand, it will increase confidence to the non-fully autonomous people to use public transports knowing that their family can "be with them". Surveillance using sensors such as cameras (cameras of different technologies can be used so that passengers' privacy is protected) and microphones, as well as smart software in the shuttle will maximize the feeling of security and the actual level of security.

Several concerns of the end users regarding the Safety and Robustness of the autonomous vehicles that are directly linked to the final User Acceptance of the new technology, can be identified. The prospective passengers fear several possible instances that could arise in case there is no driver in the bus. Indicatively:

- Passengers feeling discomfort travelling alone during nighttime
- Parents not being able to know if their kids have reached their destination safely
- Caregivers not being able to track passengers with dementia or other health issues

To address the aforementioned concerns on social and personal safety and security into the vehicle, certain measures need to be implemented. For example, third parties monitoring the route of minors or passengers with health issues could make their route much easier and less frightening. This may be followed by appropriate notifications and/or instructions to the third party, while the vehicle may also implement respective actions.

Moreover, implementing a solution for monitoring the routes of kids and patients will support safekeeping not only the users of the autonomous public shuttle but also the vehicle itself. In this section, the implementation of a video, depth and audio analytics software module for an embedded security subsystem or for cloud-based services of the system are described along with appropriate planning for the deployment and test of the service into the pilot sites of the AVENUE project.

5.1.1 Use case

The service and scenario proposes a full-fledged solution that allows designated "guardians" to follow the APT journeys of more vulnerable people, since the guardians can check the trip via a dashboard or mobile app, receive notifications via mobile app, add people to their "guarded" list, and share trips/position and E.T.A. with others.

In the context of AVENUE project, the following use cases have been identified to be further examined and addressed:

Use Case 1:

- Travelling without a guardian during nighttime can be unsettling for a vulnerable person.
- The video cameras installed in the autonomous shuttle acquire the color depth images and the data are fed into the system's video analytics algorithms for further analysis.
- When the vehicle's system identifies the passenger, the tracking begins.

Use Case 2: Kids Monitoring

- Parents need to be able to track their kids.
- The video cameras installed in the autonomous shuttle acquire the color depth images and the data are fed into the system's video analytics algorithms for further analysis.
- When the shuttle's system identifies the kid, the tracking begins, and the parents can monitor their route.

Use Case 3: Patients Monitoring

- Caregivers need to be able to track their patients, especially when they are not able to commute on their own.
- The video cameras installed in the autonomous shuttle acquire the color depth images and the data are fed into the system's video analytics algorithms for further analysis.
- When the autonomous bus's system identifies the patient, the tracking begins, and the caregivers can monitor their route.

5.2 Stakeholders (development/prototyping team)

For the "Follow my kid/grandparents" service the stakeholders are the same as in Stakeholders (development/prototyping team) (Section 2.4.2).

5.3 Technical requirements

In order to develop this service, we implement a facial recognition system capable of identifying or verifying a person from a video frame. There are multiple methods in which facial recognition systems work, but in general, they work by comparing selected facial features from a given image with faces within a database. It is also described as a Biometric Artificial Intelligence based application that can uniquely identify a person by analyzing patterns based on the person's facial textures and shape. Many face recognition techniques require multiple data of the subject in the training dataset, in order to correctly identify the face of a person. From our perspective this is not possible as we rely on one single input image of the subject. To be able to overcome this problem, we are using one shot (or single shot) facial recognition algorithms.

5.3.1 Technology

In the following sections, the research, involving algorithms and experiments conducted, is presented for the "Follow my kid/grandparents" service. As depicted in Figure 61, the first layer of sensors connects to the Hardware Abstraction Layer (HAL). The HAL implements the IP and the USB protocol supporting IP and USB cameras respectively but also can request raw data by the API endpoints in order to perform face recognition. The input data is converted and transformed in a compatible format and passed into the analytics algorithms. The prediction is then transferred via the API endpoints into the cloud. The user has access to the data and acts accordingly.

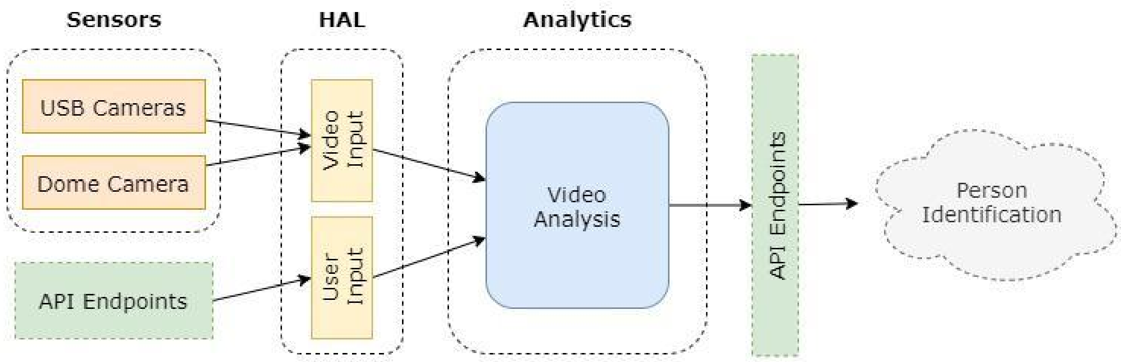


Figure 61. High level overview of the "Follow my kid/grandparents" service

One-shot learning is a classification task where one, or a few, examples are used to classify many new examples in the future. This characterizes tasks seen in the field of face recognition, such as face identification and face verification, where people must be classified correctly with different facial expressions, lighting conditions, accessories, and hairstyles given one or a few template photos.

Modern face recognition systems approach the problem of one-shot learning via face recognition by learning a rich low-dimensional feature representation, called a face embedding, that can be calculated for faces easily and compared for verification and identification tasks. Historically, embeddings were learned for one-shot learning problems using a Siamese network (Figure 62). The training of Siamese networks with comparative loss functions resulted in better performance, later leading to the triplet loss function used in the FaceNet[52] system by Google that achieved state-of-the-art results on benchmark face recognition tasks.

[52] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

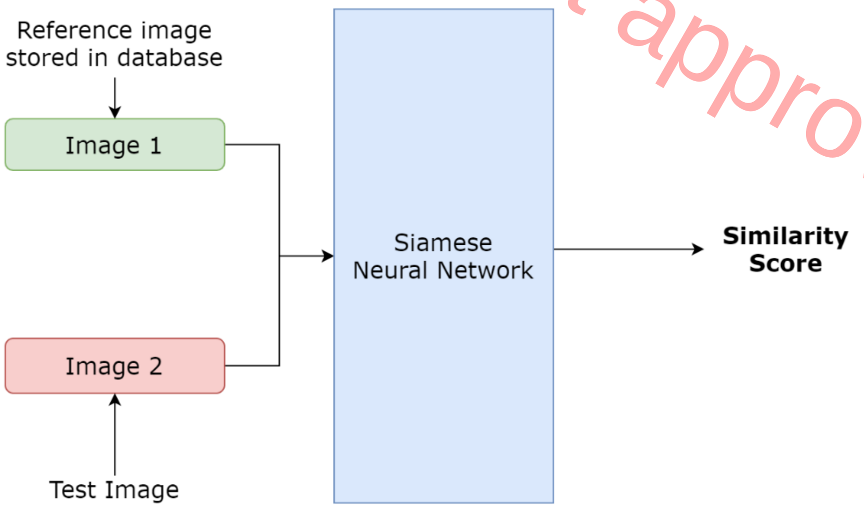


Figure 62. Architecture of a Siamese Neural Network

Instead of directly classifying an input (test) image to one of the 10 people in the shuttle, this network instead takes an extra reference image of the person as input and will produce a similarity score denoting the chances that the two input images belong to the same person. Typically, the similarity score is squished between 0 and 1 using a sigmoid function; wherein 0 denotes no similarity and 1 denotes full similarity. Any number between 0 and 1 is interpreted accordingly. Notice that this network is not learning to classify an image directly to any of the output classes. Rather, it is learning a similarity function, which takes two images as input and expresses how similar they are.

A new passenger can be enrolled to the Follow My Kid service using a single image of his face which will be stored in a database. Using this as the reference image, the network will calculate the similarity for any new instance presented to it. Thus, we conclude that the network can predict the score in one shot.

Methodology

Data collection and preprocessing: We used the MS1M-ArcFace[53] dataset for training our network and the LFW[54] for the testing. The two datasets were augmented via pre-processing to include face masks using the MaskTheFace tool[55]. For the sake of simplicity, we name the synthetic datasets MS1M-ArcFace+M and LFW+M, respectively.

Network pipeline: As current face verification models use fully connected layers, spatial information is lost along with the ability to understand the convolution features in a human sense. To address this obstacle, the plug-in xCos module is integrated as described below and presented in Figure 63.

[53] Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87{102. Springer (2016)

[54] Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)

[55] Anwar, A., Raychowdhury, A.: Masked face recognition for secure authentication (2020)

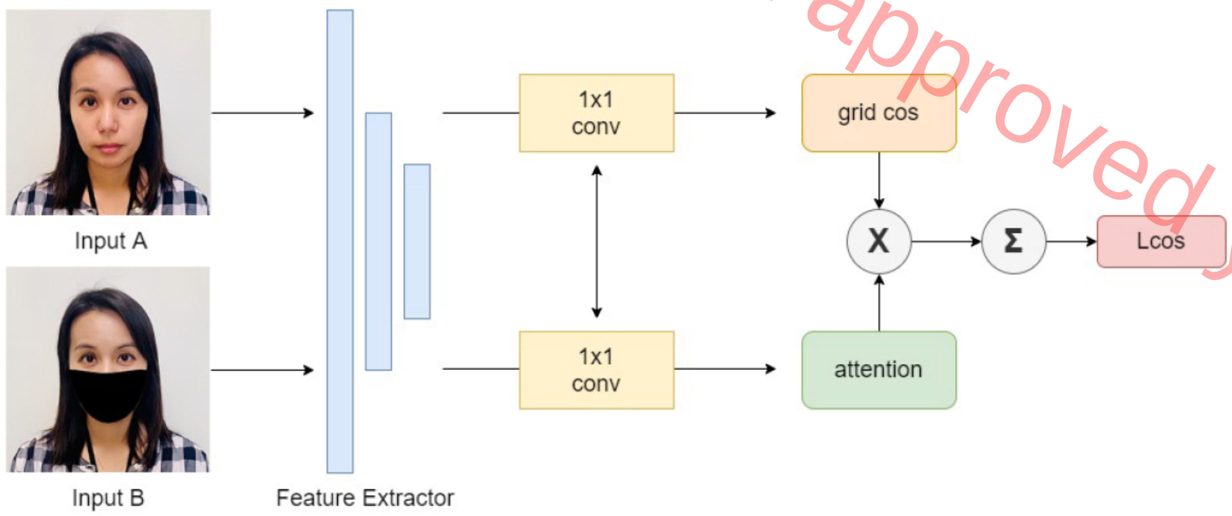


Figure 63. Network pipeline: the two input images are preprocessed and passed into the backbone CNN for feature extraction along with the plug-in xCos module.

Input: The two input images are preprocessed and passed into the feature extractor. The Input A is the database image while the Input B is the image cropped from the video stream.

Backbone: We implement the same CNN feature extractor as in ArcFace[56]. However, in order to employ the xCos module, the last fully connected layer and the previous flatten layer are replaced with an 1×1 convolutional layer.

Lcos calculation (xCos): Patch-wise cosine similarity is multiplied by the attention maps and then summed to calculate the Lcos.

5.3.2 Equipment

The same equipment is utilised that is enlisted in Section 2.4.3.2 and 2.4.3 of the Security trust service without the need for any audio sensing equipment (such as microphones).

5.4 Prototyping plan

The prototyping plan of this service consists of the development phase and the deployment phase:

Development phase: In this phase, CERTH initially plans and analyses the requirements with inputs from all the stakeholders. Once the requirement analysis is done, the final software and hardware representation and documentation of the requirements are accepted from the project stakeholders. The next stage consists of designing the critical service components, the research and the development of the algorithms and the integration of the components across several platforms. Also, in this stage several data captures are performed for training the employed machine learning algorithms. CERTH will perform additional data capture sessions with different lighting conditions, angles and operators. The same data from the "Enhance the sense of security and trust" service's Prototyping plan can be reused. An initial setup of the equipment and the camera sensors is required, along with the interconnection of the

[56] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690{4699 (2019)

components and software preparation. The captured data should demonstrate multiple camera angles and various capturing conditions. The next steps include the re-evaluation of the results, using the new data acquired for retraining the machine learning algorithms.

Deployment phase: In the deployment phase, the individual service components are unified into a complete system. The system has discrete input and output and is ready to be integrated with the other platforms. In the next months, CERTH will deploy the service in a demo AV, perform internal tests and verifications with all the stakeholders. In this stage, minor modifications and fine-tuning may be applied on the final setup, mainly on design or algorithms of the service depending on the real operation conditions.

Prototyping phase: Amobility will in coordination with CERTH install cameras and sensors in the Amobility shuttles for testing and validation of the technologies, use cases etc.

5.5 Result analysis

At the final step, an evaluation of the service under real conditions follows, that can be performed on daily or controlled routes of Amobility operator along with a safety officer inside the shuttle, depending on the GDPR permissions. Also, short-sessions for assessment with real passengers will be available. The results will also be cross validated with manual person identification by Amoblity for validation purposes of the service's accuracy and compared to the mobile application (by MT) if possible. After the successful evaluation of the service in Amobility sites, the results will be used to fine tune the service if required and the service will be deployed and evaluated also in other sites of the involved operators in the AVENUE project towards a successful integration to the relevant pilot sites.

More specifically, Table 11 shows the testing accuracy on the original MS1M-ArcFace, LFW and the artificially created MS1M-ArcFace+M and LFW+M, which contain additional masked samples. As we can see, despite the minimal loss caused by the explainability module on the original datasets, a significant improvement of about 6.2% has been accomplished in our augmented datasets which contain face masks.

Table 11. Accuracy comparison between methods and different datasets indicate a significant improvement of about 6.2% in our augmented datasets which contain face masks.

Method	***Training dataset***	***Testing dataset***	***Masks***	***Accuracy***
ArcFace	MS1M-ArcFace	LFW	No	99.83%
ArcFace-xCos	MS1M-ArcFace	LFW	No	99.35%
ArcFace+M	MS1M-ArcFace+M	LFW+M	Yes	68.33%
Ours, ArcFace-xCos+M	MS1M-ArcFace+M	LFW+M	Yes	74.52%

Figure 64 also highlights the improvements made over the original ArcFace+M. The maps are generated (a) by the original ArcFace+M while (b) by our improved model. The cosine similarity between the two faces is negative both in (a) and (b) on the bottom half of the faces as the mask portrays different characteristics such as shape and color. However, the attention map in our improved (b) model indicated that the network focuses more on the upper half characteristics around the eyes and the nose.

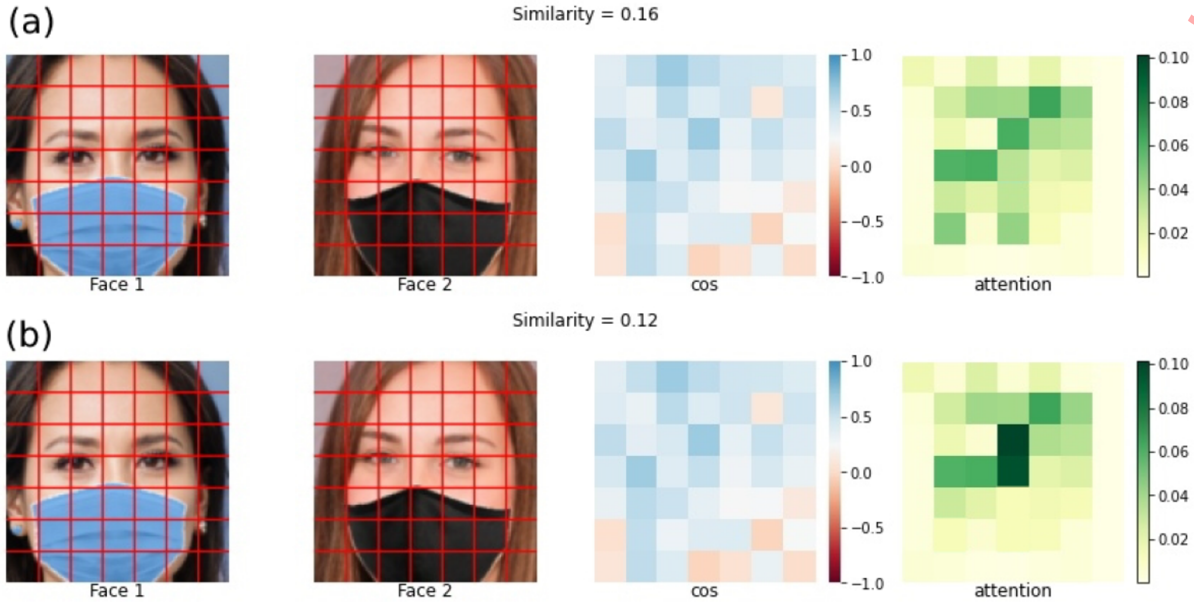


Figure 64. Input and maps generated by (a) the original ArcFace+M and (b) our improved model. Although the cosine (cos) similarity is negative both in (a) and (b) on the bottom half of the faces, the attention map in our improved (b) model indicated that the network focuses more on the upper half characteristics around the eyes and the nose, ignoring the region covered by the mask.

Figure 65 showcases an experimental real-world scenario on an autonomous shuttle, running on the NVIDIA Jetson AGX Xavier[57]. The proposed system is able to correctly identify the two passengers among a database of 20 people.

[57] NVIDIA Jetson AGX Xavier. https://developer.nvidia.com/embedded/jetson-agx-xavier-developer-kit, accessed: 2021-03-28

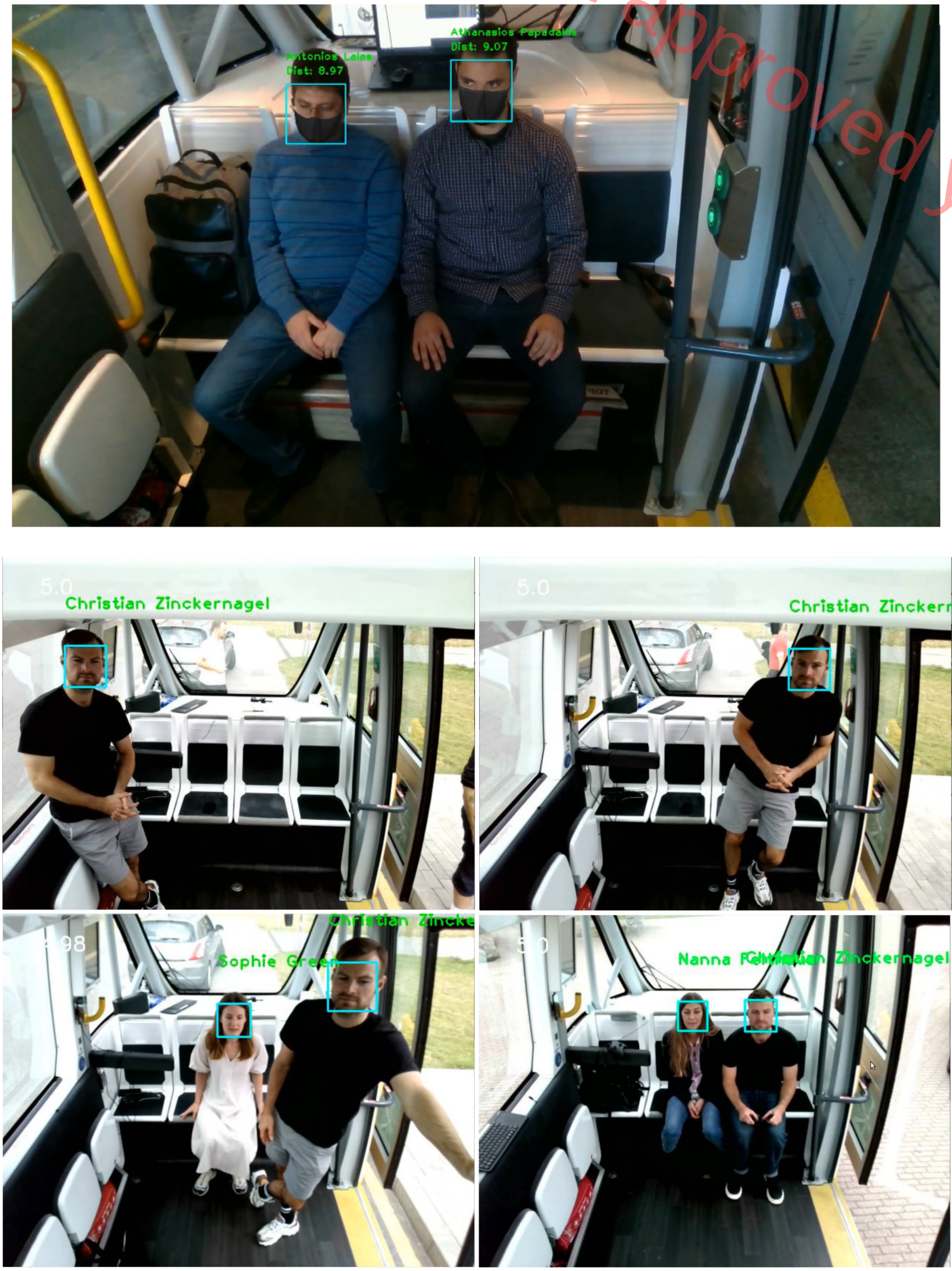


Figure 65. An experimental real-world demonstration running on the NVIDIA Jetson AGX Xavier. The proposed system is able to correctly identify the passengers of the autonomous shuttle, among a database of 10 people.

Figure 66 shows additional results of our improved model. The attention maps verify that the bottom parts of the face are efficiently ignored while the upper face characteristics have the most significant impact on calculating the similarity score.

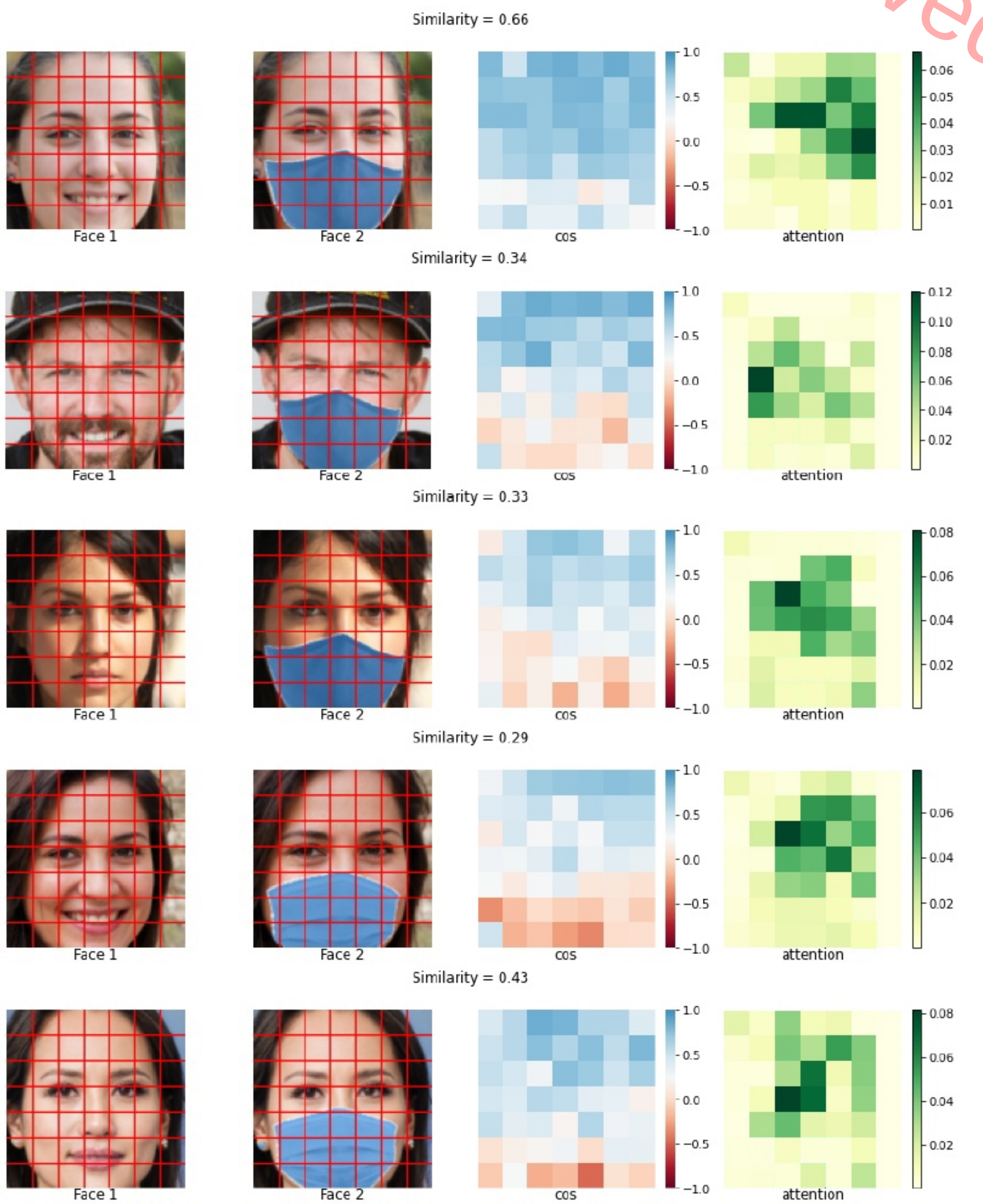


Figure 66. Additional results of our improved model. Facial images are artificially generated using StyleGAN[58] and post-processed to include masks using the MaskTheFace[59].

The system was designed as a flexible and end-to-end service accessible by an Android mobile application. In Figure 67, the architecture of the proposed system is depicted. A new passenger can be

[58] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proc. CVPR (2020)
[59] Anwar, A., Raychowdhury, A.: Masked face recognition for secure authentication (2020)

enrolled to the service via the mobile application, using a single image of his face. The image will be processed and stored
temporarily in a database on the cloud platform. By using this as the reference image, the network calculates the similarity for any new instance presented to it from the shuttle. The network can predict a similarity score in one shot and inform the client via an API call and relevant notification through the mobile phone. The proposed approach can be further extended to support homeland security surveillance infrastructures in order to mitigate domestic security risks. In this context, it is envisioned to establish a highly adaptable security framework capable of leveraging the capabilities of autonomous public transport operators as well as law enforcement agencies.

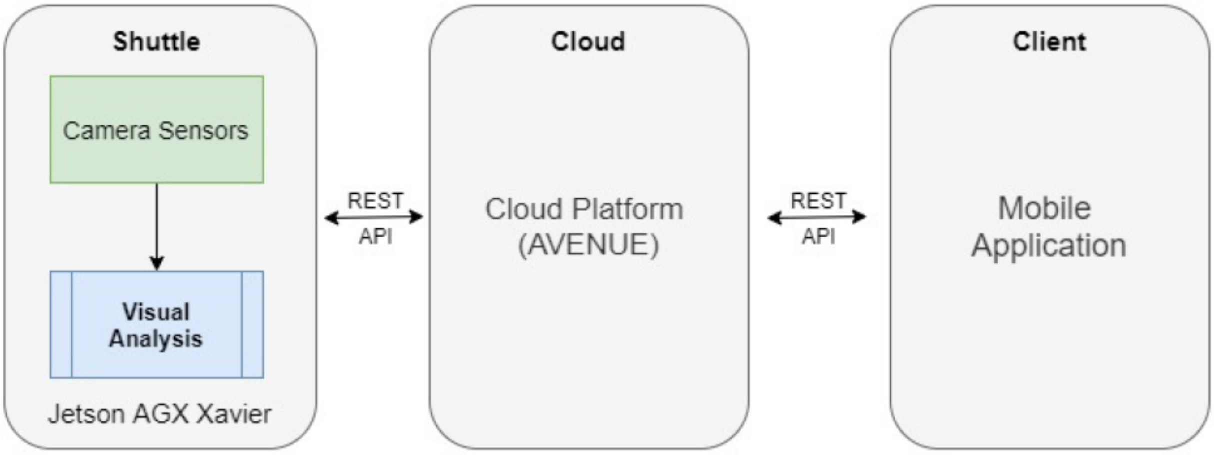


Figure 67. The architecture of the proposed system. A new passenger is enrolled by the client. A single image of his face will be processed and stored temporarily in a database and the network will calculate similarities with current passengers. The client is informed via an API call and relevant notification through the mobile phone.

6 Service: Shuttle environment assessment

6.1 Concept of service

The service "Shuttle environment assessment" aims to maintain at acceptable levels the environmental conditions in the autonomous vehicle that may not be adequately controlled due to the absence of the shuttle driver. Minimum acceptable conditions and comfort, such as good air quality, acceptable odours and absence of smoke are necessary for the safe transport of the passengers, as well as the viability of the whole autonomous service, since lack of these conditions within the vehicle could significantly discourage potential passengers. After all, monitoring the environment conditions could enable for passengers' alert and warning services via notifications, thus enhancing the user experience and safety during their trips.

Under these circumstances, there are several instances that were considered in order for the prospective passengers to feel content and safe. In particular, while there would be no driver inside the vehicle, variable problems might come up, such as the following:

- There will be no presence of the stuff inside the bus to prevent someone from lighting a cigarette

- In quite high or low temperatures, there will be no driver in order to regulate the air conditioning system
- In emergency situations there will be no one in charge of informing the operators and the competent authorities
- If the air concentration of CO2 inside the bus is high, and someone might get dizzy or exhibit breathing difficulties, there will be no driver so as to either open the windows or stop the shuttle

As far as it is considered, the buses should pose a comfortable environment for all the passengers. This feeling could undoubtedly be strengthened by controlling the temperature inside the vehicle. Besides, heating, ventilation and air conditioning control, now belong to the standard equipment on city buses. As a result, it is crucial for temperature sensors to be positioned in suitable locations inside the vehicle. Simultaneously, these sensors provide feedback to the operator who has access to the air conditioning system and in cases that the temperature exceeds a suitable limit, it can be put into operation. In that way, the existing temperature can be adjusted, in order to provide the appropriate indoor climate, namely not to be neither too hot nor too cold for the passengers.

In addition, detection of certain pollutants, such as CO2, NO2, or dust particles in the indoor environment, along with critical temperature variations, is important for the condition of certain passengers, especially ill people, such as asthma patients. Smoke in the vehicle, i.e. from a person that lights a cigarette, will deteriorate the passenger experience but may also put in danger the whole vehicle (danger of fire). Detection of certain events (air quality deterioration, smoke) raises a notification or an alert to the passengers along with instructions on how to handle this situation, while the vehicle may also implement respective actions. Likewise, smoke in the vehicle will deteriorate the passenger experience, put in danger the whole vehicle (danger of fire) and may also result in cancelling the autonomous transport service. Detection of certain events (air quality deterioration, smoke) raises a notification or an alert to the supervisor and/or the suitable authorities (i.e. police, fire department). This is followed by appropriate notifications and/or instructions to the passengers, while the vehicle may also implement respective actions.

6.1.1 Use case

The service is responsible for the timely, accurate, robust and automatic detection of any change in the air quality and the presence of smoke or fire, inside the vehicle. In such cases there is an alert on the system, notifications and instructions are sent to the passengers to the operators and/or to the suitable authorities.

In light of this, several possible situations could take place, counting from high level CO2 concentrations at the indoors air composition to presence of humidity, smoke or even fire. Exposure to carbon dioxide can produce a variety of health effects. These may include headaches, dizziness, restlessness, a tingling or pins or needles feeling, difficulty breathing, sweating, tiredness, and increased heart rate. Because people exhale CO2, in crowded places it can also mean bad ventilation which could increase the risk of virus transmission. Furthermore, detection of certain air quality indices and pollutants in the indoor environment, along with critical temperature variations are necessary for providing a secure service to the passengers. In particular, a gas composition sensor is used for checking the inside air quality. This sensor monitors the air intake of the heating, ventilation, and air conditioning system of the vehicle, while it detects undesirable gases. The operator can adjust the system accordingly, by shutting off the intake and recirculating the indoor air back to the outside. Furthermore, humidity and temperature sensors are used for measuring/regulating indoor air quality and adjusting the heating, ventilation, and air conditioning system settings accordingly.

Summarizing, it is passengers' wish to travel in a clean and comfortable environment and be notified if the conditions have deteriorated. Moreover, operators would like to be notified when environmental conditions are considered harmful for the passengers. For a deeper explanation, some of the most representative use cases are indicatively displayed as following:

Use Case 1: Lighting a Cigarette

- Inside the autonomous vehicle a passenger lights a cigarette.
- The smoke detection systems embedded inside the vehicle detects the smoke coming out of the cigarette
- The real-time sensor data is sent to a central PC, installed in the vehicle and in which the data processing take place
- With the real-time data processing the PC decides that it is an emergency case and sends the message of smoke detection to the operators
- The suitable operator evaluates the criticality of the situation and decides how to intervene (with an announcement from the loudspeakers or by taking more drastically measurements, for example by stopping the bus).

Use Case 2: Exposure to Carbon Dioxide

- While the bus is on its route, high levels of Carbon Dioxide are detected from the relevant sensor.
- The sensor data are sent in real time to the central PC installed in the vehicle and are processed.
- Carbon Dioxide in unusual levels of air concentrations might have an adverse effect on passengers' health. For instance, high levels of CO_2 are considered to be related with dizziness, restlessness or breathing difficulties and increased heart rate. It can also mean bad ventilation which increases the virus transmission risk.
- In order to prevent an event concerning these health issues from taking place, such as a passenger's fainting, the PC is sending to the vehicle's operator a message via the dashboard. The operator can decide to open the windows, so as the air comes back to its normal composition. Simultaneously, the passengers are informed about the air composition through the corresponding mobile application in real-time.

Use Case 3: High Temperature on the Autonomous Vehicle

- It is a hot day in summer and indoors the temperature is starting to rise.
- The suitable sensor measures the temperature and sends the data to the central PC installed in the vehicle.
- When the temperature exceeds a predefined level, the PC sends to the vehicle's central system the command to put the cooling system into operation.
- At the same time, passengers are able to be informed of the temperature inside the vehicle through the mobile application.

6.2 Stakeholders (development/prototyping team)

In this section, the relevant stakeholders involved in the development and prototyping of the Shuttle Environment Assessment Service are introduced. In particular:

- **CERTH** is conducting innovative research regarding the development of the environmental assessment in terms of setting up the sensors inside the vehicle. Simultaneously, CERTH will determine the sensors that will be needed, the specific role that every single sensor has to play and how this equipment will be efficiently organized and combined so as the operators will have access to the crucial information collected from sensors.
- **Bestmile** provides multiple integration interfaces towards the different stakeholders in the project (e.g., vehicle manufacturers, public transport operators) into its cloud platform. In this service, Bestmile provides connectivity between the sensors' system and the related operators regarding the notifications that are generated by the detection software.
- **MobileThinking** is responsible for the AVENUE mobile application development. This application will illustrate the environment-related measurements or some other relevant with the passenger's convenience information, in a user-friendly environment, while it will provide any necessary notifications to the passengers. MobileThinking is also responsible for the development of monitoring dashboard software. The role of the dashboard is to collect the information sensed in vehicles, to determine the critical events in vehicles and to notify the operators or the intervention teams.
- **Amobility** is the operator of the autonomous vehicles in the Copenhagen and Oslo site and is handling all daily operation of the vehicles and everyday contact with the end users. As an operator Amobility provides the autonomous shuttles that will be used for the AVENUE pilot activities and its facilities for performing data capture activities, deploying the service and testing its performance through short or longer evaluation periods.

6.3 Technical requirements

The service is based on the collection of several vehicle telemetry information from sensors and related services. Accordingly, several sensors are required, such as:

- Environmental sensors (for assessing pollution, air quality in the vehicle, such as CO2, NO2, dust particles concentrations, temperature in the vehicle, humidity sensors, fogging prevention sensors etc.)
- Smoke sensors and smoking detection (for fire and also people that smoke in the vehicle)

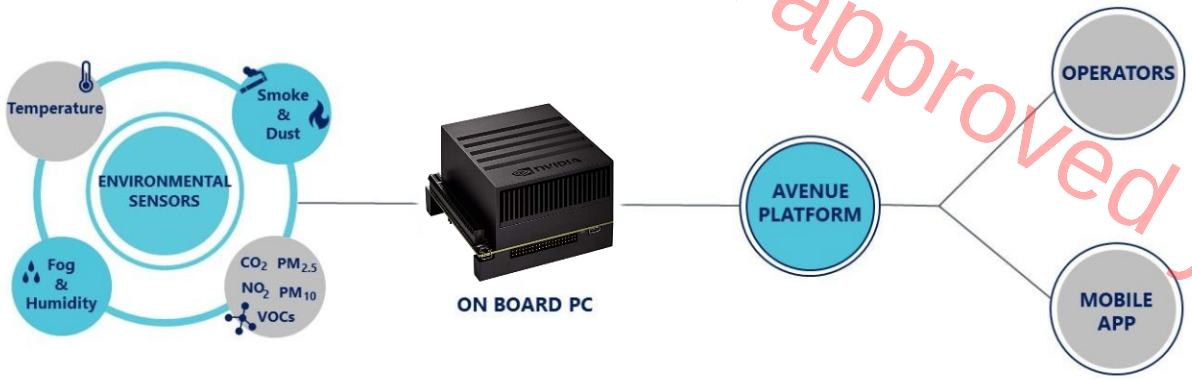


Figure 68. Environmental Assessment Process

Obviously, a wide variety of sensors should be embedded into the vehicle for this purpose, along with proper connectivity with a central assessment system. Subsequently, the conditions inside the bus will be monitored in order to ensure stability. In more detail, collection of the sensor data, data processing and subsequently, data transmission to a PC and furthermore to operators seems to be essential so as for the condition stability inside the vehicle to be guaranteed. Likewise, it is necessary, an interface with the central AVENUE platform to be developed, as well as, a mobile application so that passengers will be able to have access to information regarding the status of the vehicle or other relevant information. As a result of such implementation, when it is required, notifications to the passengers and supervisors will be provided and in case of alarms, the supervisors or other suitable authorities, will be notified in order to deal with the situation or prevent an unfortunate event. A summary of the entire environmental assessment process is depicted in Figure 68.

6.3.1 Technology

The technology utilised is based on the different sensors that will be used to develop and deploy this service. A detailed explanation of the technology of each sensor is out of the scope of this deliverable. Indicative sensors are presented in the next section.

6.3.2 Equipment

Overall, multiple environmental sensors will be needed, so as for the quality of the air to be determined, as well as sensors about smoke detection or fogging prevention. Moreover, there must be appropriate action to assure the power supply to the sensors.

In particular, regarding the sensors that will be used, some indicative models are proposed as it follows:

Aeroqual

- Series 500 – Portable Indoor Air Quality Monitor: It is compatible with 30 different sensors. It has the ability of real-time data observation through a screen, while all the metrics are stored in memory and can be transferred to PC via a USB. [60]

[60] Aeroqual, Series 500 – Portable Indoor Air Quality Monitor, Available online at: https://www.aeroqual.com/product/series-500-portable-indoor-monitor

- Indoor Air Quality Test Kit (Starter): It measures PM2.5, PM10, CO2, VOC, temperature and humidity, as it uses the suitable sensor. Also, it can connect to a PC through a USB and the measurements are captured on a screen. [61]

TSI

- AIRASSURE PM2.5-AD INDOOR AIR QUALITY MONITOR IPM2.5-AD: It can measure PM2.5. It is placed onto the wall and it includes a screen, in which the measurement is illustrated[62]
- Q-TRAK INDOOR AIR QUALITY MONITOR 7575: It has the ability of measuring CO2, CO, temperature and humidity. On its screen it could be able to capture up to 5 measurements, while VOC sensors could be connected. Furthermore, it can store up to 39-days data, with samples of 1-minuite frequent. Communication through Bluetooth is also available.[63]

IOTSENS

- IOTSENS city: It can measure PM1, PM2.5, PM10, temperature and humidity. It is a complete solution, as it has the ability to connect through LoRaWAN, Sigfox, NBIoT, GPRS, Wifi, while there are specific platforms and applications from the company.[64]

Senseair

- EXPLORAPM2.5: It can measure PM2.5, temperature and humidity. It connects through LoRaWAN and there is a specific cloud-based platform and application.[65]

- EXPLORACO2: It is able to measure CO2, temperature and humidity. It connects through LoRaWAN and there is a specific cloud-based platform and application. In addition, it has the ability to connect with an application, developed from externals, using the open API.[66]

- SENSEAIR AERCAST: It is able to measure CO2, temperature and humidity. It connects through BLE as well as other choices such as LoRa, ZigBee, και NBIoT.[67]

Awair

---

[61] Aeroqual, Indoor Air Quality Test Kit (Starter), Available online at: https://www.aeroqual.com/product/indoor-portable-monitor-pro-kit

[62] Aeroqual, AIRASSURE PM2.5-AD INDOOR AIR QUALITY MONITOR IPM2.5-AD, Available online at: https://tsi.com/products/indoor-air-quality-meters-instruments/installed-iaq/airassure-pm2-5-ad-indoor-air-quality-monitor-ipm2-5-ad/

[63] Aeroqual, Q-TRAK INDOOR AIR QUALITY MONITOR 7575, Available online at: https://tsi.com/products/indoor-air-quality-meters-instruments/indoor-air-quality-meters/q-trak-indoor-air-quality-monitor-7575/

[64] IOTSENS, IOTSENS city, Available online at: http://www.iotsens.com/solution/smart-city/

[65] Senseair, EXPLORAPM2.5, Available online at: https://senseair.com/products/aercast-explora-family/explorapm2-5/

[66] Senseair, EXPLORACO2, Available online at: https://senseair.com/products/aercast-explora-family/exploraco2/

[67] Senseair, SENSEAIR AERCAST, Available online at: https://senseair.com/products/aercast-explora-family/aercast/

- Awair Element: It is able to measure CO2, VOCs, PM2.5, temperature and humidity. It connects through Bluetooth and includes an application for the control of the measurements. Additionally, a LED is available to indicate dimensions.[68]

- Awair Glow C: It is able to measure VOCs, temperature and humidity. It is connected to the socket and devices such as a fan, dehumidifier, heater can be connected to it, which are activated when the limit we have set for a measurement is exceeded. It has an application and has the ability to connect to Google Home.[69]

- Awair 2nd Edition: It is able to measure VOCs, CO2, PM2.5, temperature and humidity. It has an application for controlling the measurements and can be connected to Google Home, Alexa and Ecobee. It has a Led to indicate the measurements.[70]

- Awair Omni: It has the ability to measure VOCs, CO2, PM2.5, ambient light, ambient noise, temperature and humidity. It has a Led to indicate the measurements. It can connect to WiFi, Bluetooth, Cellular, Ethernet. It is mounted on the wall.[71]

Airthings

- Wave Plus: It has the ability to measure radon, CO2, VOCs, pressure, temperature and humidity. It connects via Bluetooth and has its own application and platform for displaying and processing data.[72]

- Wave Mini: It has the ability to measure VOCs, temperature and humidity. It has Led for visual display and is connected via Bluetooth. It can be used in conjunction with Google Assistant and has its own application.[73]

Netatmo

- Netatmo: It has the ability to measure air quality, noise, temperature and humidity. Moreover, it is able to connect via Bluetooth and has its own application.[74]

---

[68] Awair, Awair Element, Available online at: https://getawair.com/pages/awair-element

[69] Awair, Awair Glow C, Available online at: https://getawair.com/pages/awair-glow

[70] Awair, Awair 2nd Edition, Available online at: https://getawair.com/pages/awair-2nd-edition

[71] Awair, Awair Element, Available online at: https://getawair.com/pages/awair-for-business

[72] Airthings, Wave Plus, Available online at: https://www.airthings.com/wave-plus

[73] Airthings, Wave Mini, Available online at: https://www.airthings.com/wave-mini

[74] Netatmo, Available online at: https://www.netatmo.com/en-us/aircare/homecoach

Kaiterra

- Laser Egg: It has the ability to measure PM2.5, temperature and humidity. It has WiFi connectivity, its own application and can connect to other smart devices via Apple Home Kit. It has a screen on which the measurements are illustrated, as well as the weather forecast.[75]

- Laser Egg + Chemical: It has the ability to measure PM2.5, VOCs, temperature and humidity. It has WiFi connectivity, has its own application and can connect to other smart devices via Apple HomeKit. It has a screen on which the measurements are illustrated, as well as the weather forecast. [76]

- Laser Egg + CO2: It has the ability to measure PM2.5, CO2, temperature and humidity. It has WiFi connectivity, has its own application and can connect to other smart devices via Apple HomeKit. It has a screen on which the measurements are illustrated, as well as the weather forecast.[77]

6.3.3 Methodology

Smoke can be detected with the parameters measured by air quality sensors. As it has been demonstrated by many researchers, these parameters can be used to train machine learning algorithms to detect smoke[78]. To this direction, the problem can be handled as a classification problem. Two common classification algorithms are KNNs and SVMs.

- *KNN:* As demonstrated in Figure 69, it works by finding the distances between a point of interest and all the points in the data using Euclidean distance. The specified number of points (K) closest to the point of interest is selected and then the most frequent label is voted as the class that represents the point of interest.

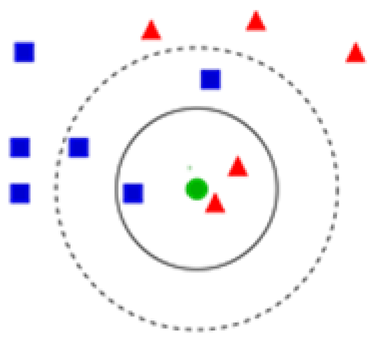


Figure 69: Example of KNN classification with two classes

[75] Kaiterra, Laser Egg, Available online at: https://www.kaiterra.com/en/laser-egg/

[76] Kaiterra, Laser Egg + Chemical, Available online at: europe.kaiterra.com/collections/frontpage/products/laser-egg-plus-chemical

[77] Kaiterra, Laser Egg + CO2, Available online at: https://www.kaiterra.com/en/laser-egg-co2

[78] J. H. Cho, "Detection of Smoking in Indoor Environment Using Machine Learning," *Applied Sciences,* vol. 10, p. 8912, 2020.

- *SVM:* As depicted in Figure 70, SVMs achieve a representation of the data in a higher dimensional feature space through the kernel "trick". There the data become linearly separated and decision boundaries can be set to distinguish between classes.

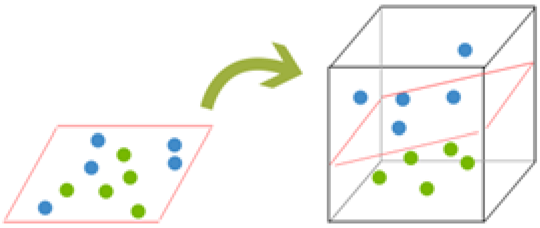


Figure 70 Example of SVM classification in a two-dimensional dataset

Regarding CO2 monitoring, the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), in its Informative Appendix D to *ANSI/ASHRAE Standard 62.1-2016: Ventilation for Acceptable Indoor Air Quality*[79] notes that indoor CO2 concentrations no greater than 700 parts per million (ppm) above outdoor CO2 concentrations will satisfy a substantial majority (about 80%) of occupants. Since outdoor CO2 concentrations typically range between 375 to 500 ppm, a threshold of 1000 ppm is a logical choice for alerting the operator and the users that air quality conditions inside the shuttle might be deteriorating. Nevertheless, it should be noted that if at the time of CO2 measurement the respective area is not occupied at its usual density, it is not an effective indicator of ventilation. High CO2 concentrations could mean that other indoor pollutants may also be increased. In this case, the amount of outdoor air introduced into the respective space may need to be increased.

6.4 Prototyping plan

The prototyping plan of this service consists of the development phase and the deployment phase:

Development phase: In this phase, CERTH initially plans and analyses the requirements with inputs from all the stakeholders. Once the requirement analysis is done, the final software and hardware representation and documentation of the requirements are accepted from the project stakeholders. The next stage consists of designing the critical service components, the research and the development of the algorithms and the integration of the components across several platforms. At this point, all the individual sensors that will be used, were specifically determined and either built or bought. Following, they were connected all together at the AVENUE Platform.

Specifically, the Awair element sensor was determined to be the most suitable solutions for handling the aforementioned use cases. It's installation and connectivity was tested in CERTH premises. For setting up, Awair element is connected via Bluetooth through Awair's phone app. Once the Bluetooth connection is established it can be connected to Wi-Fi address with the phone and then the sensor can communicate with Awair's servers. Even though it is possible to access Awair's data through the servers, by default there is a daily request limit and possible latency. A more feasible solution is to enable the local API

[79] https://www.ashrae.org/technical-resources/standards-and-guidelines/read-only-versions-of-ashrae-standards

feature[80] which can stream the data via a localhost. A local feed that updates every 10 seconds is generated and it can be accessed through Awair's IP address. Finally a python script to request data from this feed every 30 seconds was developed and embedded to the system. Those values are then streamed into the dashboard.

On top of that, awair's potential to detect smoke was investigated. To this direction, an extensive data capture was performed. The capture included cigarette smoking for a couple of hours and with a 10 second sampling interval. This experiment was performed in a controlled space of approximately 20 square metres, where the temperature and humidity was controlled by an air conditioner and different ventilating options were simulated by opening one window, two windows or no window at all. The subject lighting the cigarettes performed this process repeatedly and in different locations of the room so that the capturing is not biased. Also, background measurements without smoking and under the same conditions were performed because they are required to train the smoke detection classifiers.

After data pre-processing it was concluded that Awair's parameters that are highly affected by smoke are $PM_{2.5}$ and TVOC, meaning that they could be useful input variables for the classification models to learn to distinguish between smoking and regular conditions. This is well depicted in Figure 71, where the boxplot distributions for smoking and not smoking, for both of these variables are provided. It is apparent that both the range, as well as the median, are shifted while smoking.

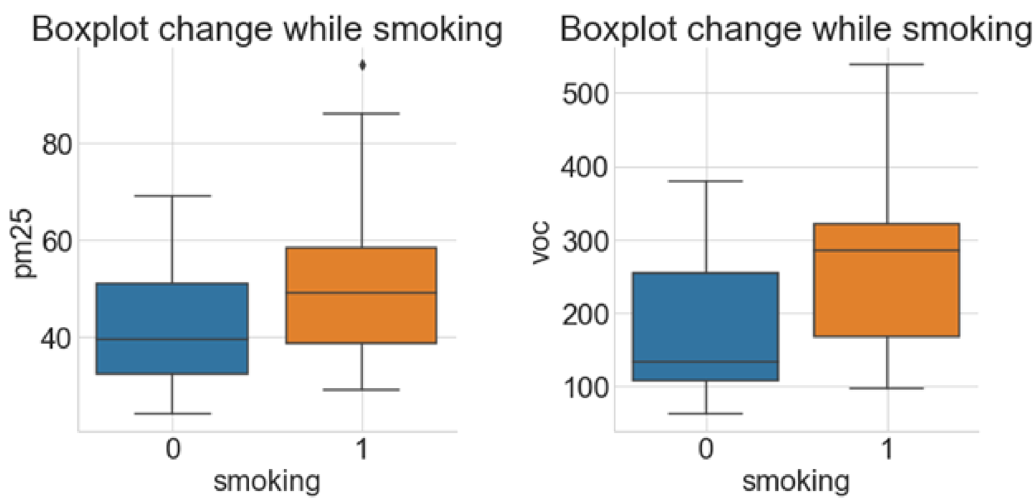


Figure 71: Boxplot shift between smoking and not smoking for both $PM_{2.5}$ and TVOC

Considering this, an SVM model was trained on these data using $PM_{2.5}$ and TVOC as input to detect smoke. The model was deployed in the overall system, inputs from the local API are fed to it every minute and a prediction for smoking or not smoking is generated and displayed via the dashboard.

Deployment phase: In the deployment phase, the individual service components are unified into a complete system. The system has discrete input and output and is ready to be integrated with the other platforms. In the previous months, CERTH deployed the Awair element sensor and the service in a demo AV, performed internal tests and verifications with all the stakeholders. Subsequently, the sensor was connected with the PC installed into the vehicle (NVIDIA Jetson AGX Xavier), as well as with the AVENUE platform and the mobile application. Then, for a sufficient period, the installed sensors measured the corresponding metrics such as CO2, NO2 humidity, temperature, dust, smoke, etc. After data processing,

[80] https://support.getawair.com/hc/en-us/articles/360049221014-Awair-Element-Local-API-Feature

and through a calibration process, it was made possible for the conclusions we are interested in to be extracted. Based on the obtained information, the rates of successful measurements was determined. Minor modifications and fine-tuning was applied on the final setup, mainly on design and algorithms of the service depending on the real operation conditions.

Prototyping phase: Amobility, in coordination with CERTH installed cameras and sensors in the Amobility shuttles for testing and validation of the technologies, use cases etc.

6.5 Result analysis

At the final step, the evaluation of the whole system, from the sensor measurement process to the appropriate messages and measurements appeared onto the users' mobile application, took place under real conditions, which can be performed on daily or controlled routes of Amoblity operator along with a safety officer inside the shuttle. Also, short-sessions for assessment with real passengers were available. The results were also cross validated with manual person notification by Amoblity for validation purposes of the service's accuracy.
The measurements from the sensors were processed and used in order to monitor the condition indoor regarding the air quality. These measurements, as well as the conclusions of the real-time assessment, were sent both to passengers and operators. After the successful evaluation of the service in Amobility sites, were deployed and evaluated also in other sites of the involved operators in the AVENUE project towards a successful integration to the relevant pilot sites.

More specifically, both KNNs and SVMs were trained on this data. Due to the simplicity of the data, both algorithms predicted all cases in the test dataset correctly, but the KNN algorithm displayed signs of overfitting. It achieved a 100% accuracy which should not be possible, but showed no skill when tested on new data. On the other hand, the SVM achieved a 99.3% accuracy on the first training attempt and a 99.5% accuracy after parameter grid searching. Figure 72 depicts the model's confusion matrix and a classification report is included in Table 12.

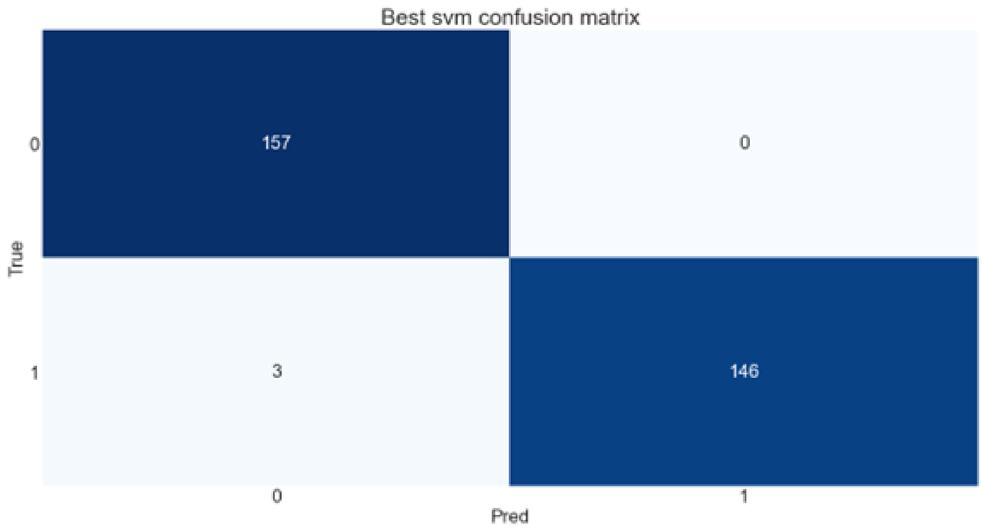


Figure 72: Confusion matrix of SVM smoking classifier

When tested on new data the model seemed to be able to distinguish between smoking and not smoking under proper ventilation conditions. Considering the above, the SVM model was deployed in the AVENUE setup taking as input the per minute average values of TVOC and $PM_{2.5}$ and generating an output for whether smoke is detected or not. One pitfall that was observed, is that when tested, the model produced lots of false alarms when ventilation conditions were poor. The original data capturing was performed in a well ventilated environment and the model translates not smoking conditions as such. In any case proper ventilation is set as a minimum viable condition that should be met at all times inside AVENUE shuttles.

Table 12: Classification report for SVM smoking model

	Precision	***Recall***	***F1-Score***	***Support***
Not smoking	0.99	1.0	1.0	157
Smoking	1.0	0.99	1.0	149
Macro Avg	1.0	1.0	1.0	306

7 Service: Smart feedback system

7.1 Concept of service

The service “Smart feedback system” aims at allowing the travellers inside the shuttle to give easy and effortless feedback to the operators of the shuttle - when the safety officer is no longer inside the shuttle. It is important for the operators to know if people are satisfied with the services and transportation. Currently the safety officer talks to the travellers and via his/hers presence also becomes the conversation channel between the operators and the travellers. When the safety officer is removed, knowing whether they are satisfied or disappointed can be even more important, as the safety officer is not there to support, hence knowing how to assist the travellers.

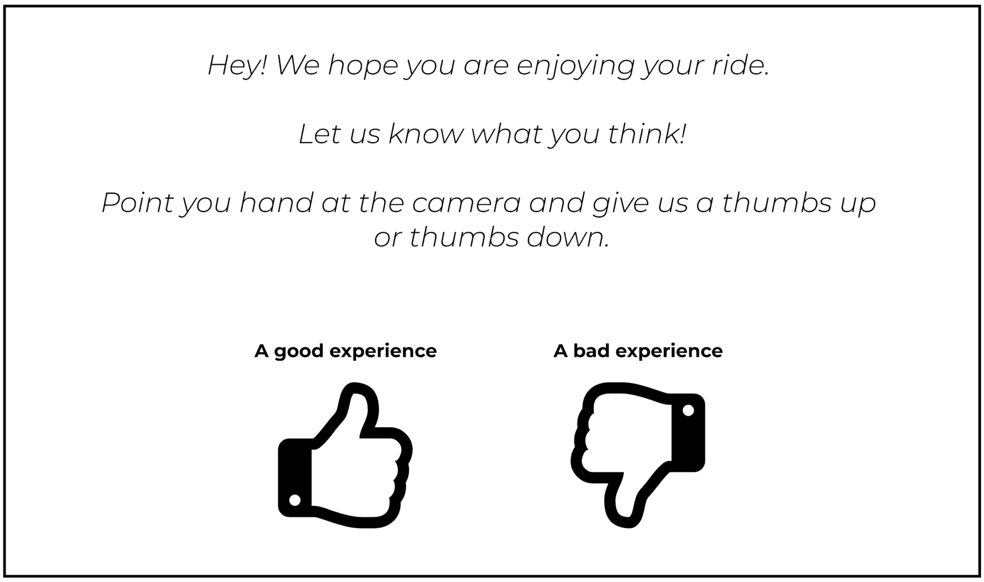
When removing the operator, automated services have to perform the same level of service and interaction as he/she did while being in the shuttle.

This service aims at allowing travellers to give their feedback about liking/disliking the service experience as easy as possible. This will be done by instructing the travellers to give a hand gesture to one of the cameras inside the shuttle. This will allow the travellers to effortlessly say “i like” or “i dont like” the experience with a thumbs up or a thumbs down.

The concept will be communicated to the travellers via stickers inside the shuttle. To begin with camera technology will be used to capture the thumbs up or thumbs down, but if possible sounds sensors will also be tested to capture the experience/feedback from the travellers.

7.1.1 Use case

The “smart feedback service” will allow travellers to give their feedback by a hand gesture to the cameras inside the shuttle. The service will be communicated to the travellers as follows:



Use case 1: Giving a thumbs up/down in light settings

- Mid day with sunlight
- Good visibility for the cameras

Use case 2: Giving a thumbs up/down in dark settings

- Early morning or night with no sunlight
- Low visibility for the cameras

Use case 3: Giving a thumbs up/down in crowded settings

- Many passengers inside the shuttle, both standing and seating
- Low visibility for cameras due to people standing close to the cameras

Use case 4: Giving a thumbs up/down in empty settings (or few passengers)

- Little or no passengers inside the shuttle
- Good visibility for the cameras, easy to see the hand gesture

7.2 Stakeholders (development/prototyping team)

For the “Smart feedback system” service the stakeholders are the same as in Stakeholders (development/prototyping team) (Section 2.4.2).

7.3 Technical requirements

Gesture recognition refers to the whole process of tracking human gestures to their representation and conversion to semantically meaningful commands. Research in hand gesture recognition aims to design and develop such systems that can identify explicit human gestures as input and process these gesture representations. In this service, we implement a hand recognition system capable of detecting hand gestures from a video frame. Although there are multiple methods for hand gesture recognition, we focus on simplicity and high performance using Single Shot Detectors and shallow CNN models for classification.

7.3.1 Technology

In the following sections, the research, involving algorithms and experiments conducted, is presented for the "Smart feedback system" service. As depicted in Figure 73, the first layer of sensors connects to the Hardware Abstraction Layer (HAL). The HAL implements the USB protocol supporting USB. The input data is converted and transformed in a compatible format and passed into the analytics algorithms. The output prediction is then transferred via the API endpoints into the cloud, available for further processing from other platforms.

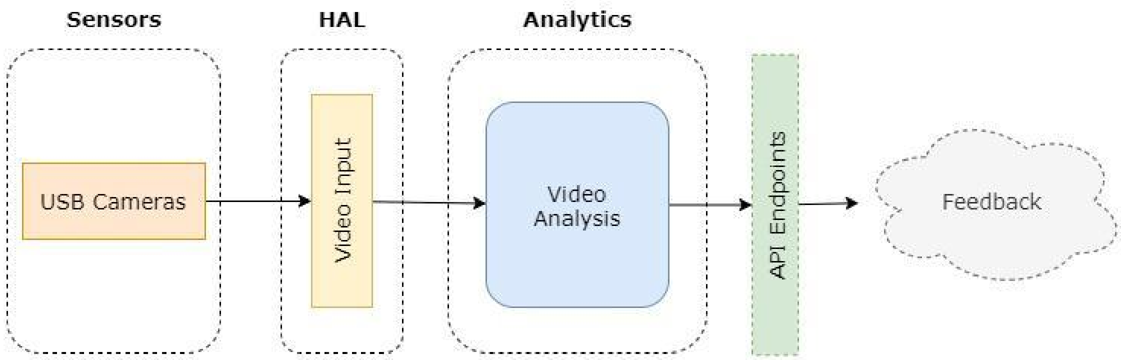


Figure 73. High level overview of the "Smart feedback" service

In the first stage, we use the Single Shot Detector (SSD) for the detection of a bounding box of where the hand(s) is and the corresponding cropped frame. The SSD approach is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detections. The early network layers are based on a standard architecture used for high quality image classification (truncated before any classification layers), which are called the base network2. Auxiliary structure is added to the network to produce detections with the following key features.

Multi-scale feature maps for detection: Convolutional feature layers to the end of the truncated base network. These layers decrease in size progressively and allow predictions of detections at multiple scales. The convolutional model for predicting detections is different for each feature layer.

Convolutional predictors for detection: Each added feature layer (or optionally an existing feature layer from the base network) can produce a fixed set of detection predictions using a set of convolutional filters. These are indicated on top of the SSD network architecture in Figure 74. For a feature layer of size m × n with p channels, the basic element for predicting parameters of a potential detection is a 3 × 3 × p small kernel that produces either a score for a category, or a shape offset relative to the default box coordinates. At each of the m × n locations where the kernel is applied, it produces an output value. The bounding box offset output values are measured relative to a default box position relative to each feature map location. Default boxes and aspect ratios. A set of default bounding boxes is associated with each feature map cell, for multiple feature maps at the top of the network. The default boxes tile the feature map in a convolutional manner, so that the position of each box relative to its corresponding cell is fixed. At each feature map cell, the offsets relative to the default box shapes in the cell are predicted, as well as the per-class scores that indicate the presence of a class instance in each of those boxes. Specifically, for each box out of k at a given location, c class scores and the 4 offsets relative to the original default box shape are computed.

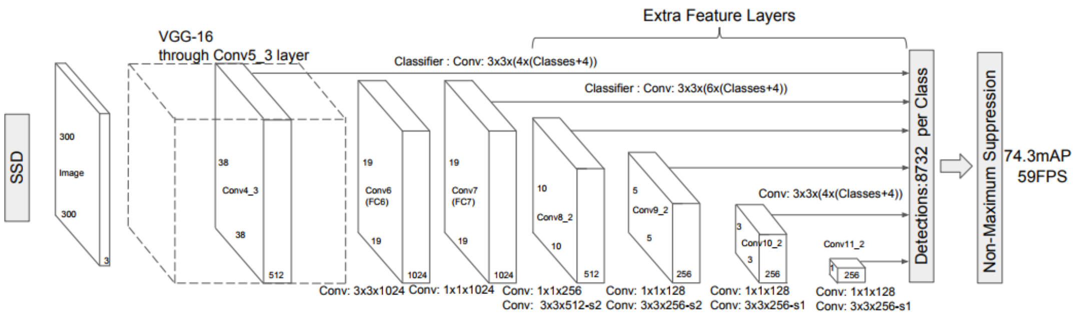


Figure 74. SSD architecture using VGG16 for feature extraction

This cropped frame of the hand is then passed to the CNN (Figure 75), which predicts a class vector output of values between 0 and 1. The values correspond to the probability of the frame to be one of the classes.

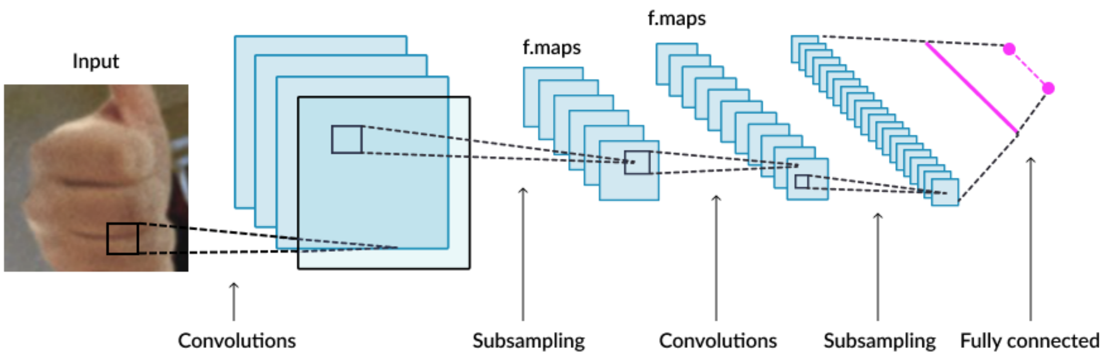


Figure 75. CNN architecture

7.3.1.1 Research results

The model is trained end-to-end and regularized so that it distils the most compact profile of the normal patterns of training data (Figure 76) and effectively detects the “thumbs-up and thumbs-down” gestures (Figure 77). The system will be extended and more gestures will be supported in the future.

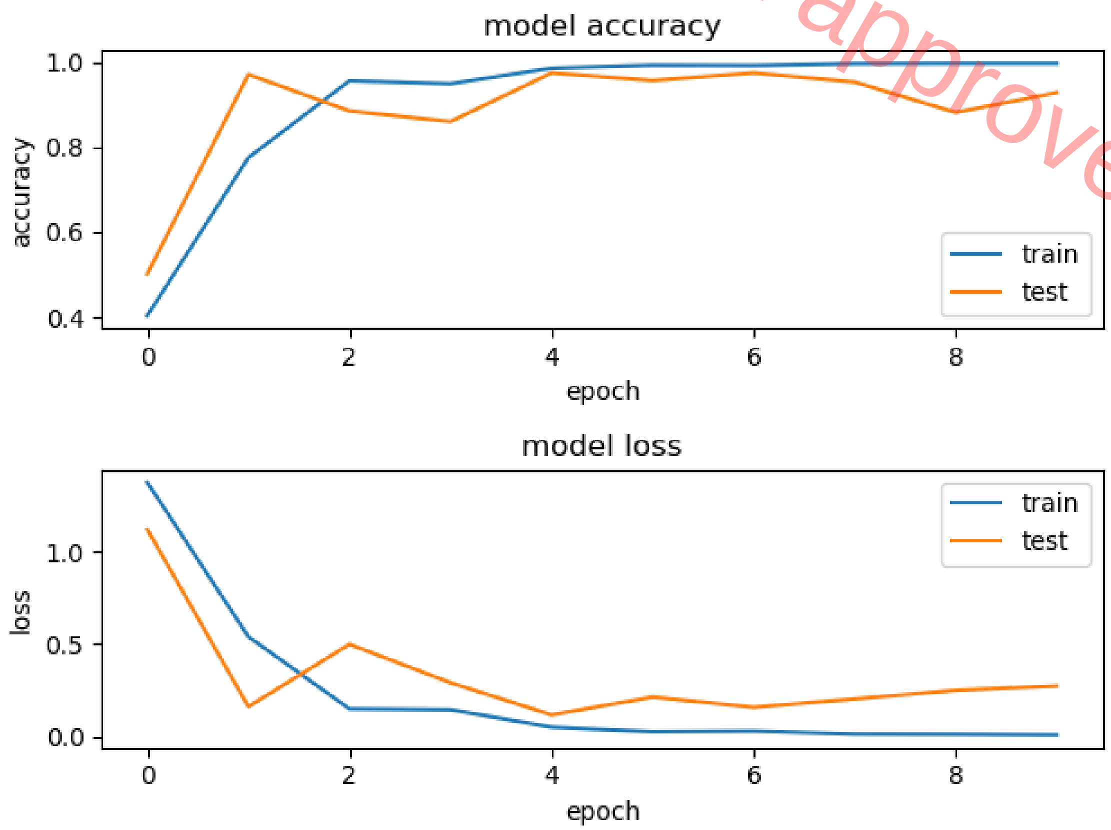


Figure 76. Training accuracy and loss at 10 epochs

Figure 77. Inference results: (a) Detection of "thumbs-up" gesture. (b) Detection of "thumbs-down" gesture

7.3.2 Equipment

The same equipment is utilised that is enlisted in Section 2.4.3.2 and 2.4.3 of the Security trust service without the need for any audio sensing equipment (such as microphones).

7.4 Prototyping plan

The prototyping plan of this service consists of the development phase and the deployment phase:

Development phase: In this phase, CERTH initially plans and analyses the requirements with inputs from all the stakeholders. Once the requirement analysis is done, the final software and hardware representation and documentation of the requirements are accepted from the project stakeholders. The next stage consists of designing the critical service components, the research and the development of the algorithms and the integration of the components across several platforms.

Deployment phase: In the deployment phase, the individual service components are unified into a complete system. The system has discrete input and output and is ready to be integrated with the other platforms. In the next months, CERTH will deploy the service in a demo AV, perform internal tests and verifications with all the stakeholders. In this stage, minor modifications and fine-tuning may be applied on the final setup, mainly on design or algorithms of the service depending on the real operation conditions.

Prototyping phase: Amobility will in coordination with CERTH install cameras and sensors in the Amobility shuttles for testing and validation of the technologies, use cases etc.

7.5 Result analysis

At the final step, an evaluation of the service under real conditions follows, was performed on daily or controlled routes of Amobility operator along with a safety officer inside the shuttle, depending on the GDPR permissions. Also, short-sessions for assessment with real passengers were available. The results were also cross validated with manual person feedback counting by the Amoblity operator and were used to fine tune the service.

Figure 78. Smart Feedback service is detecting passenger gestures in real-time

8 In-vehicle service interfaces

As a part of the in-vehicle services two interfaces have been developed to support the operation of the system and to support the operators of the system. The maintenance interface is seen as the "backend" interface for maintenance and development. And the passenger safety interface is seen as the "frontend" operator interface used during operations to monitor and navigate the shuttle environment. Both are described below.

8.1 Maintenance interface by CERTH

In the context of the AVENUE project, a maintenance interface was built, communicating all the information recorded by the sensors included in the project to a human operator. The interface is also capable of producing notifications by utilising the outputs of the AI algorithmic infrastructure. A screenshot of the dashboard is provided in Figure 79.

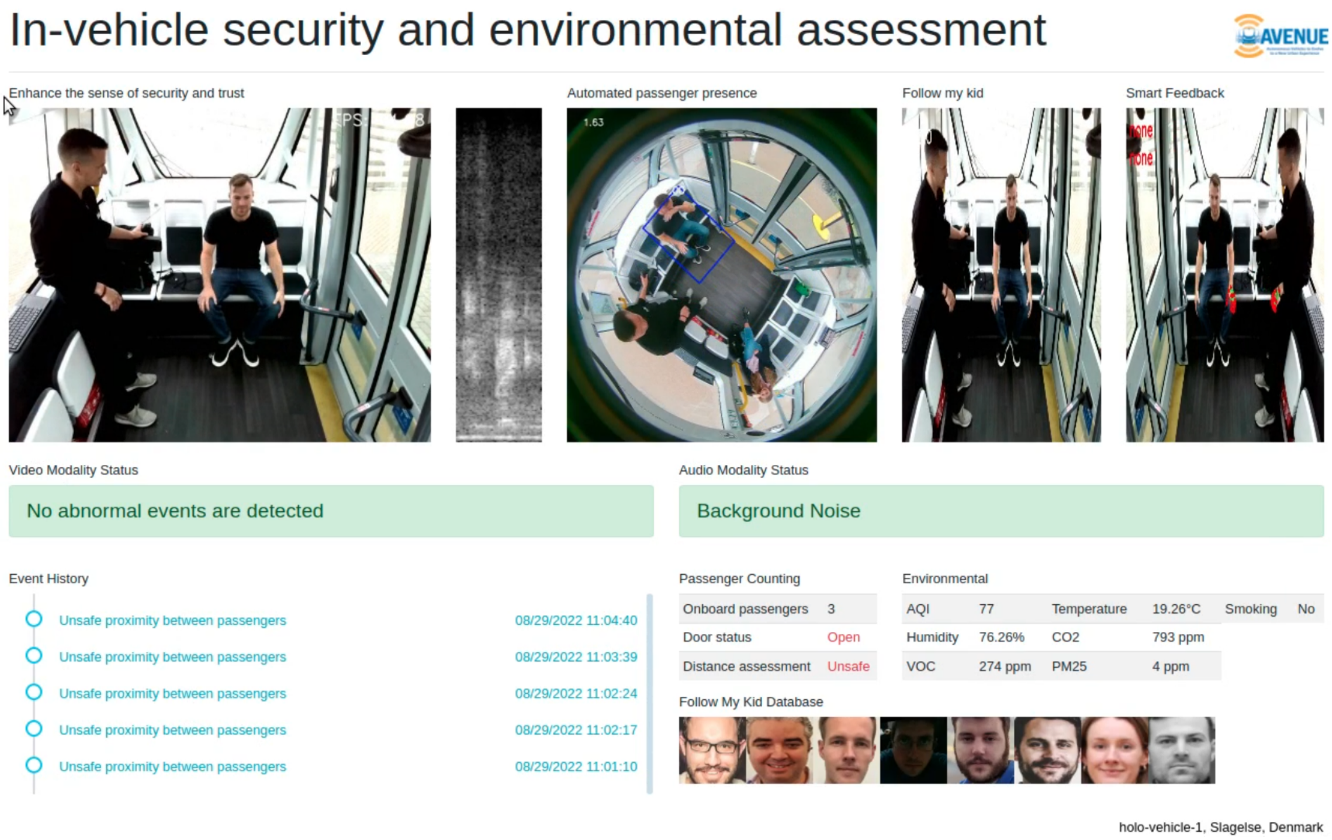


Figure 79. An overview of the dashboard displaying metrics for various services inside the autonomous shuttle ecosystem

The **Modality Status Bars** provide useful messages about the signals that each sensor type is recording. On the bottom left, the **Event History** tab provides messages informing the operator about the system's status, connectivity and the algorithms that are running in the backend. On the top panels, feeds by different modalities (RGB and acoustic) are provided with respect to the different services. In cases of abnormal events or detected objects, the displayed feed is enriched with bounding boxes and alerting messages.

The **Environmental** tab monitors all the parameters measured by the Awair element air quality sensor. Humidity and temperature information are provided which are essential for monitoring the comfort levels of indoor spaces. Also, CO_2 concentrations are displayed which can help the operator determine if

the shuttle is properly ventilated. In a crowded shuttle, if ventilation is poor CO_2 levels will rise. Finally, an estimation of harmful chemicals and fine particles (VOC and PM2.5) is provided by the sensor, the per minute average of those variables is also used as input in the smoke detection model that runs in the backend and feeds the smoking indication every minute.
The **Passenger Counting** tab, displays the outputs of AI algorithms that use camera feeds to estimate the number of passengers inside the shuttle.
Finally, the **Follow My Kid Database** tab at the bottom right, displays people in need of special attention (could be kids or elders), whose photos were stored in the database so that the system is able to track them.

8.2 Passenger safety interface by MobileTHinking

To ensure the passenger safety and comfort, the PTO will assign an employee to the role of the "Site Manager". That role is dedicated to monitoring the passenger safety and comfort in a fleet of vehicles. The monitoring is enabled by Passenger Safety Dashboard.

The dashboard enables and supports transversal services. It is a "meta service" that allows site managers to monitor the operations of other services.

The Passenger Safety Dashboard (PSD) is a layer placed on top of many other services developed within the AVENUE project, coming from in or out of vehicle service partners. Currently, it allows the PTOs' to monitoring for the following services:

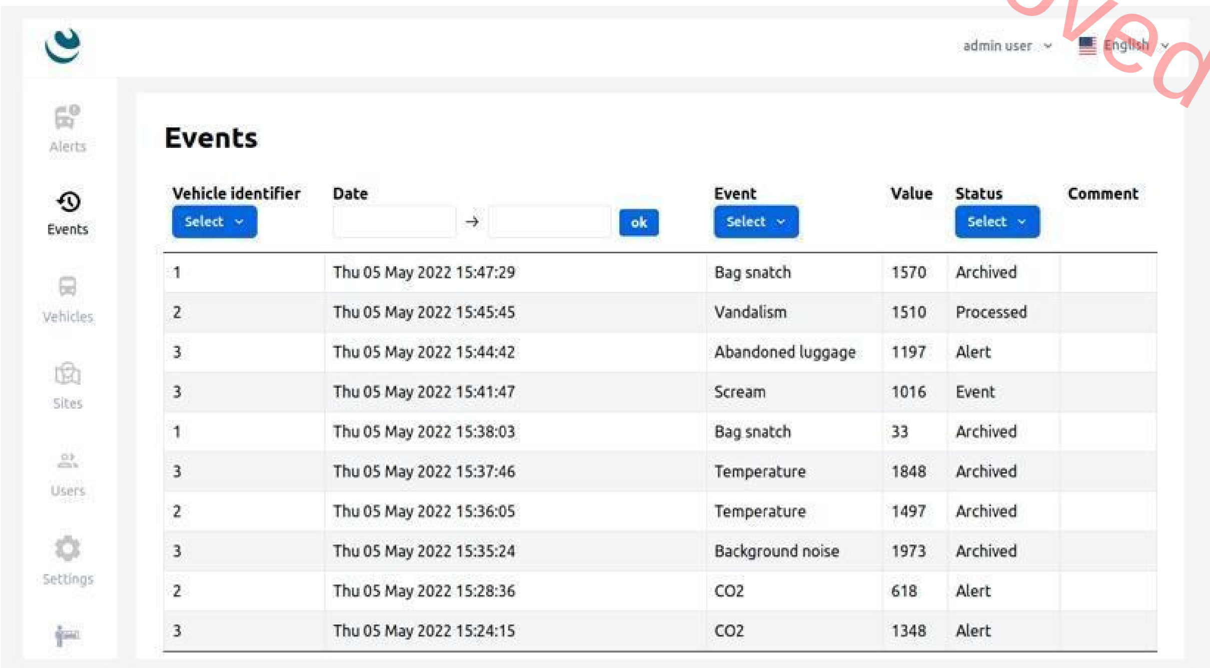
- Enhance the sense of security and trust,
- Automated passenger presence,
- Shuttle environment assessment, and
- Smart feedback system.

On top of the monitoring of existing services, it allows PTOs to :
- manage access rights per sites,
- define alerts level and settings,
- access logs of all the events that occurred,
- manually log extra information on triggered alerts

A complete work on the passenger safety dashboard has been described in the **deliverable D4.9_v1_1 or the WP 4**.

The dashboard features have been implemented according to the service requirements. They have been enhanced in several iterations.
The current main features implemented on the dashboard are:
- Event history
- Alert settings
- Alert management
- Fleet management
- Site management
- User management

Event history is a log of information collected by the in-vehicle sensors and sent to the dashboard. Events are collected when an abnormal behaviour occurs (eg. bag snatch) or as a continuous measurement (eg. temperature). This feature allows for collecting of events as well as the visualisation of their history.



Vehicle identifier	Date	Event	Value	Status	Comment
1	Thu 05 May 2022 15:47:29	Bag snatch	1570	Archived	
2	Thu 05 May 2022 15:45:45	Vandalism	1510	Processed	
3	Thu 05 May 2022 15:44:42	Abandoned luggage	1197	Alert	
3	Thu 05 May 2022 15:41:47	Scream	1016	Event	
1	Thu 05 May 2022 15:38:03	Bag snatch	33	Archived	
3	Thu 05 May 2022 15:37:46	Temperature	1848	Archived	
2	Thu 05 May 2022 15:36:05	Temperature	1497	Archived	
3	Thu 05 May 2022 15:35:24	Background noise	1973	Archived	
2	Thu 05 May 2022 15:28:36	CO2	618	Alert	
3	Thu 05 May 2022 15:24:15	CO2	1348	Alert	

Figure 80: Event history graphical interface

Alert settings are the rules that will transform an event into an alert if it exceeds a predefined limit. In fact, an event in itself is not necessarily an alerting information. For example in the context of the Automated passenger presence service, an event carries information about the number of passengers onboard. It is only when the passenger number exceeds a maximum acceptable load that the event becomes an alerting information. The alert setting feature allows PTO to define those rules that will identify the critical levels in the events and transform them into alerts.

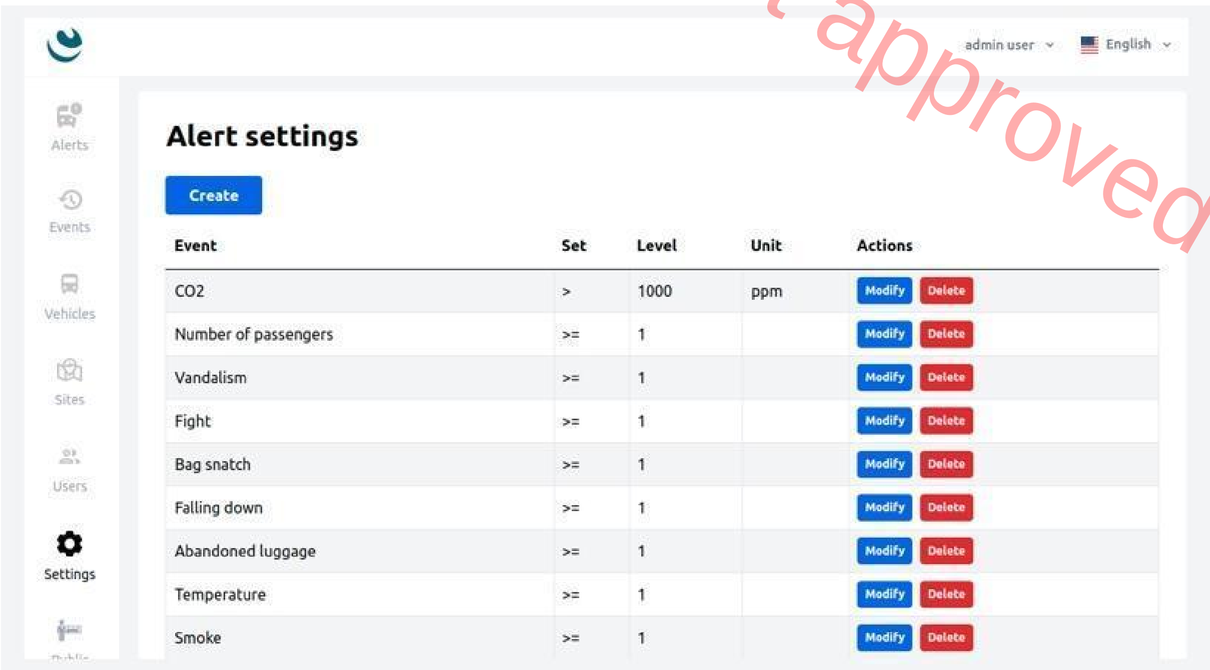


Figure 81: Alert settings graphical interface

Alert management shows alerts in a visually noticeable way. The dashboard pushes the alerts in the browser window and updates the page automatically. It allows site managers to be visually notified and to easily distinguish alerts. It also allows site managers to: quickly archive an alert with or without a comment, or to bulk archive multiple alerts by type or all alerts at once.

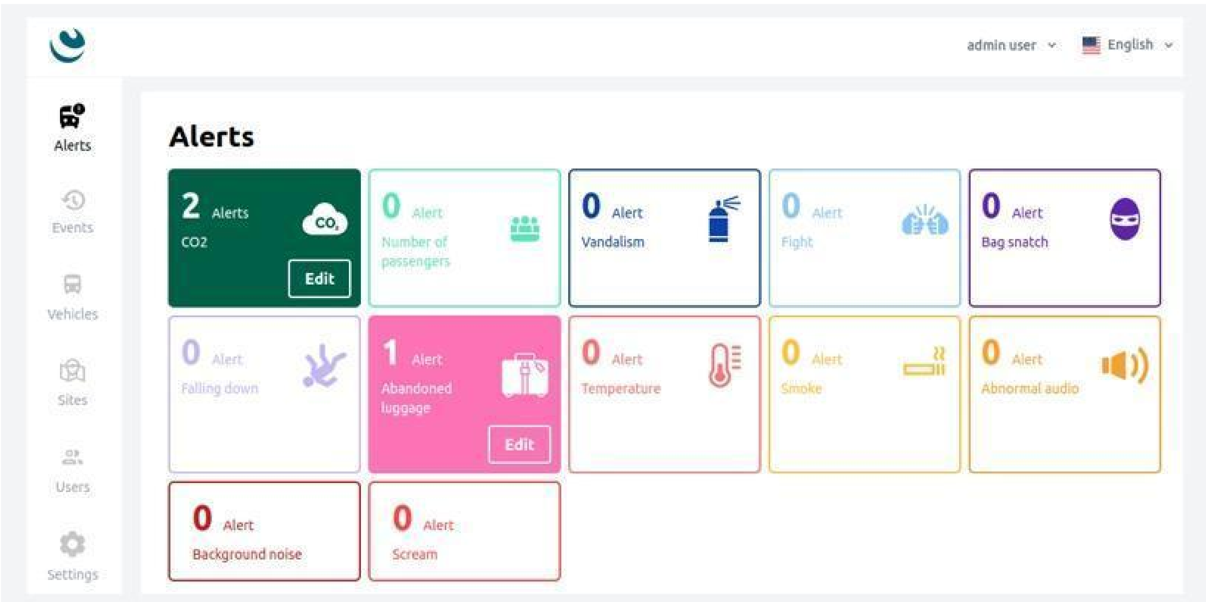


Figure 82: Alert management graphical interface

Fleet management allows PTO to register the vehicles on the dashboard. The dashboard only shows events and alerts from the registered vehicles.

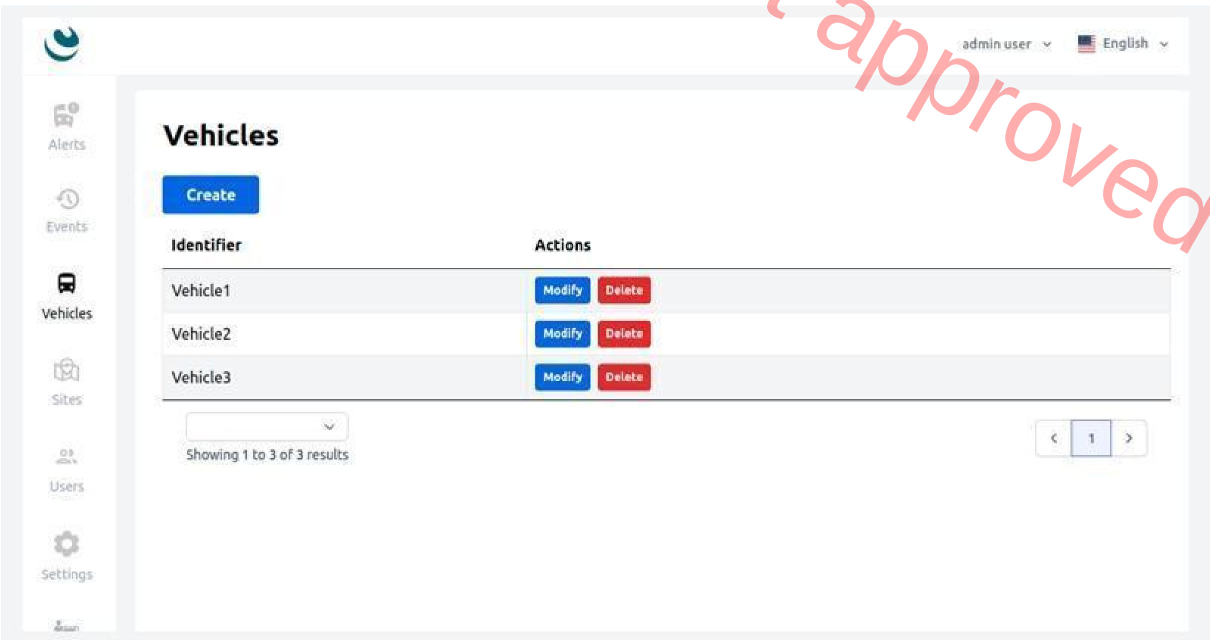


Figure 83: Fleet management graphical interface

Site management allows PTO to create a sub-fleet and to assign it to a user making him a site manager. Site management is a feature that isolates events, alerts, and vehicles per site.

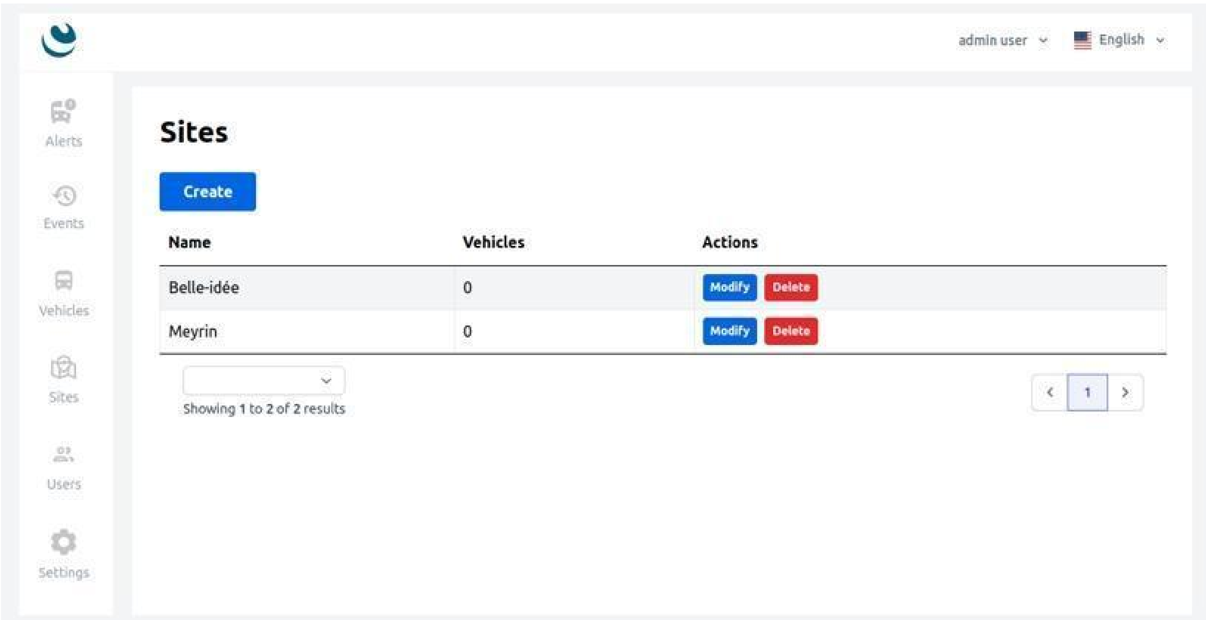


Figure 84: Site management graphical interface

User management feature allows PTO to create/update/delete users and their personal information.

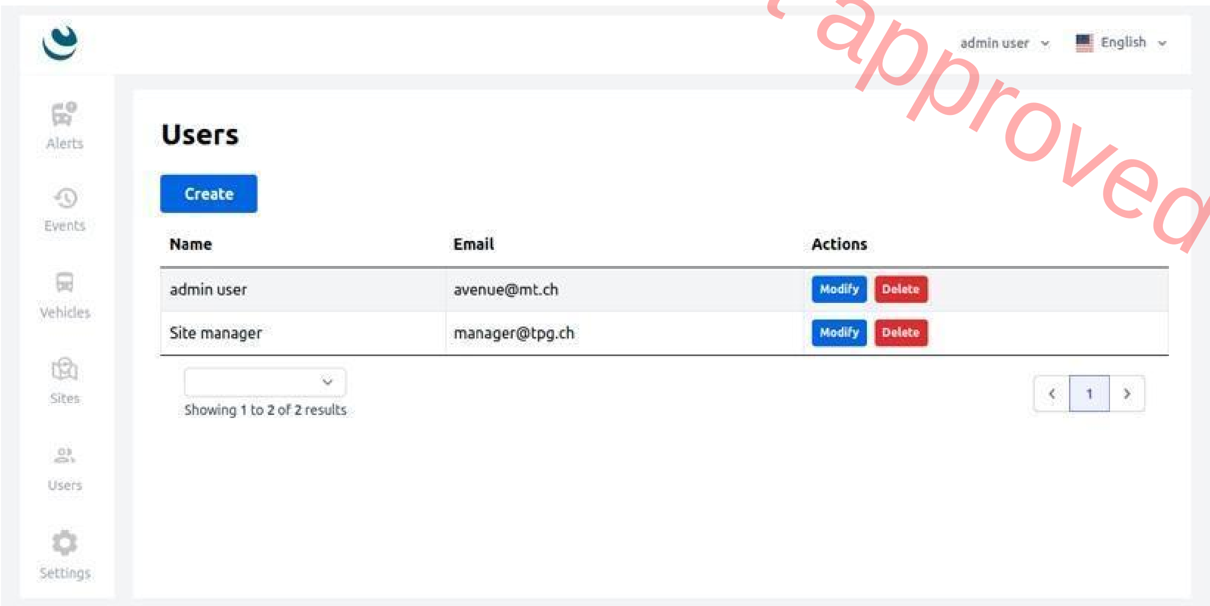


Figure 85: User management graphical interface

9 Installation guide final camera solution

Based on the learnings from the first two deliverables and a lot of testing the in-vehicle service is now developed enough to be deployed in real life shuttle environment on the AVENUE routes. To do that CERTH and Amobility as developed an installation guide enabling other partners to install and test the same camera solutions.

9.1 Equipment

Applicable for: For Navya generation 1 shuttles
Timeframe: 3-4 hours depending on mechanical capabilities
Equipment list: What you need to have before starting

- 10 " monitor + 3 m power cable (bought separately)
- HDMI cable
- Wireless keyboard for Jetson (USB wireless)
- Jetson + 3 m power cable (bought separately)
- Wifi antenna cables for jetson
- USB extension cable (3 slots)
- Inverter with two outlets

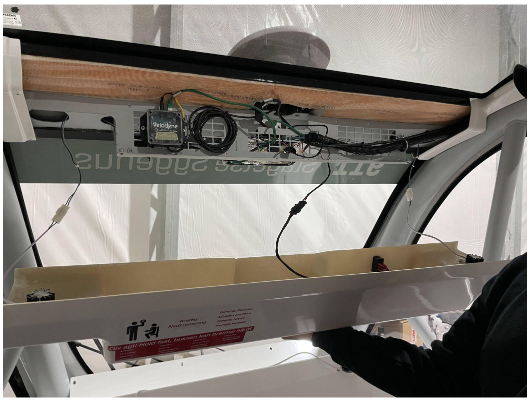
- Webcam for side camera
- Microphone
- Wide lens camera for top camera
- Ethernet cable for connecting jetson and top camera
- 3 meter cable for power to the top camera
    - Merged with power cable from Monitor 12V
- USB-A 2.0 han / MICRO-B han - 3 m for power to the microphone

9.2 Guide

To install the Jetson computer with sensors and power cables, it is necessary to take the shuttle a bit apart. This way cables and the jetson can be hidden away nicely and the whole setup will look more professional. Ideally all panels are removed before mounting the equipment and then tested before the panels are put back on. As a part of the process a lot of screws will be removed, so remember where they are placed or write it down where they come from to avoid putting it wrongly back together. Installing the equipment has 26 steps as follows in table 13.

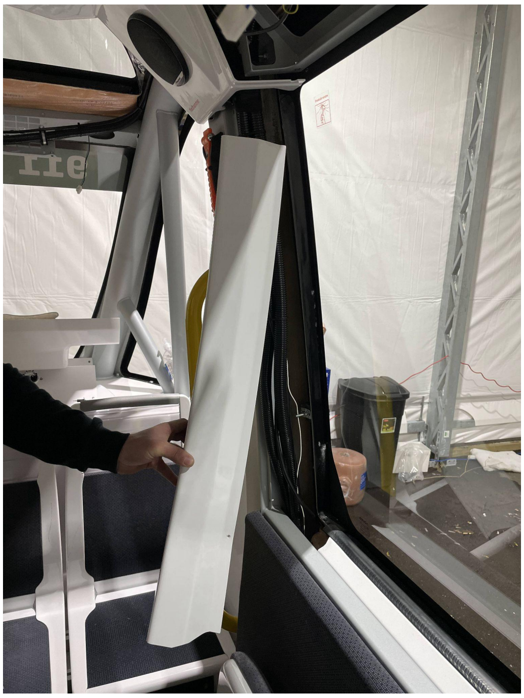
Table 13: Installation guide

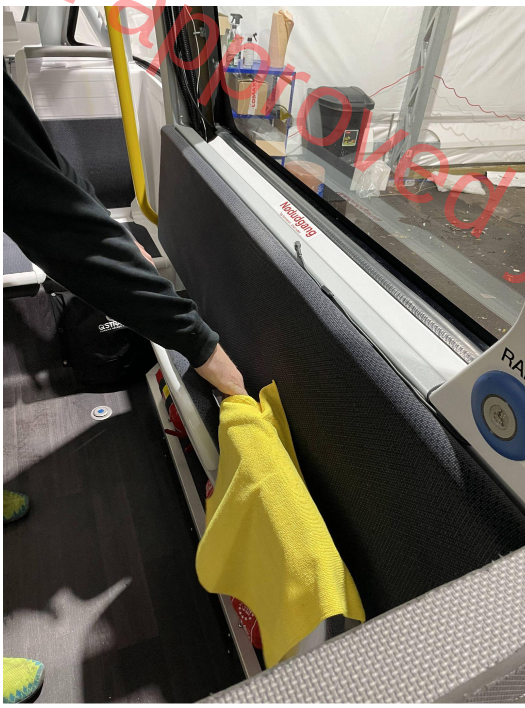
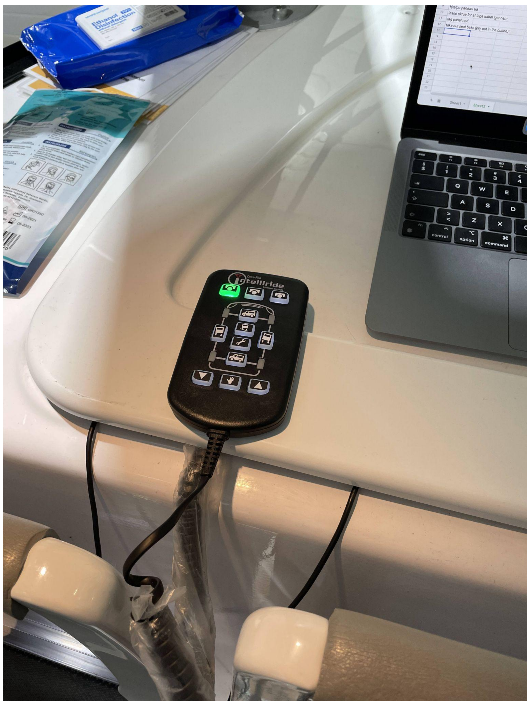
1	Remove the large top panel over the side window of the shuttle, by taking out the scews. Remember to unplug the light cabels attached to the inside of the panel.	

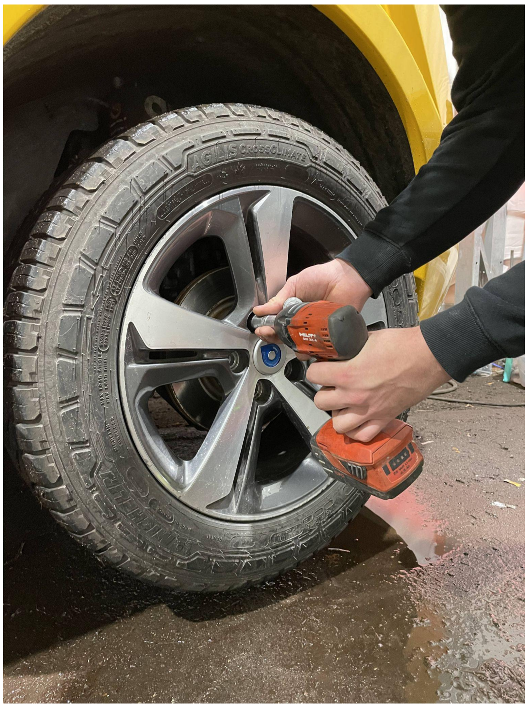
2	Remove the screws in one end of the metalplate holding the to tu pannels together - closest to the window, and loosen the other screw a bit, so that the metalplate can be twisted to the side, allosing you to wire the cables later between the two top panels.	
3	Remove the large top panel in the left side of the shuttle, by taking out the scews. Remember to unplug the light cabels attached to the inside of the panel.	 

		
4	Remove the next panel to the right above the smaller side window, you can take it out slowly by pulling the mid panel towards the ceiling gently.	 

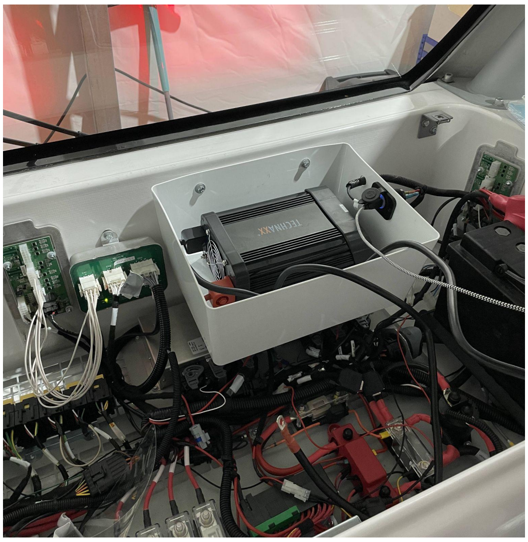
5	Remove the screw in the top of the vertical panel with the speaker, to ensure that the cables can be wired behind the panel.	 

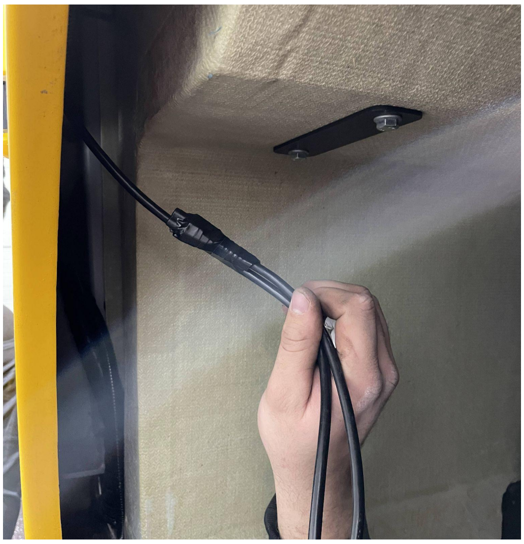
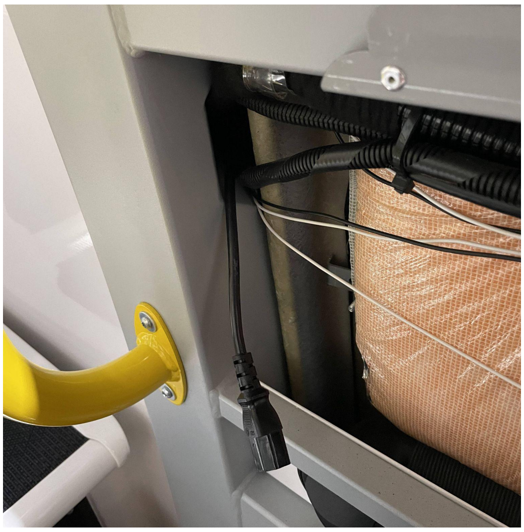
6	Remove the vertical panel alongside the window towards the foldable seats. Drag the panbel downwards to get it out easier.	 

7	Remove the seat-back of the foddable seats by pulling out the board in the buttom.	 
8	Take vehicle to highest suspension position. Once done, turn of the vehicle.	

9	Go outside and prepare to take off the rear left wheel. Use a pump jack and a safety to make sure the vehicle stays up. You need to work under the wheelarch. One up, remove the bolts and take of the wheel.	 
10	Inside the wheelarch, remove the plastic cover, by first disconnecting the air suspension (little black wire for airflow), then removing the plastic fittings inside the wheelarch. Once done you should be able to see wires coming out of the battery department.	

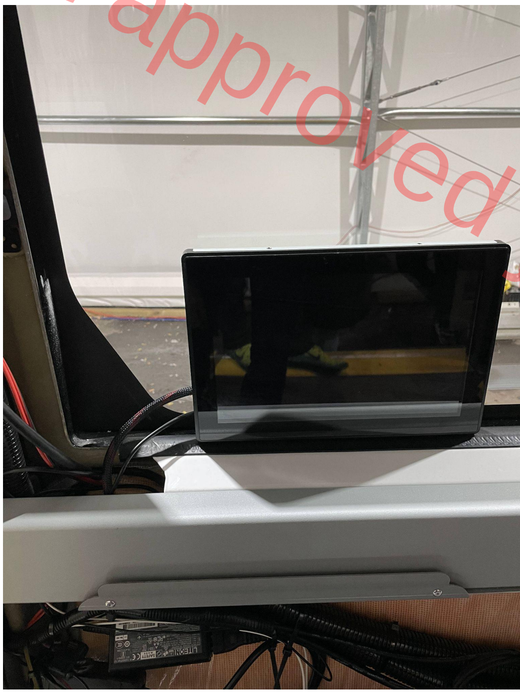
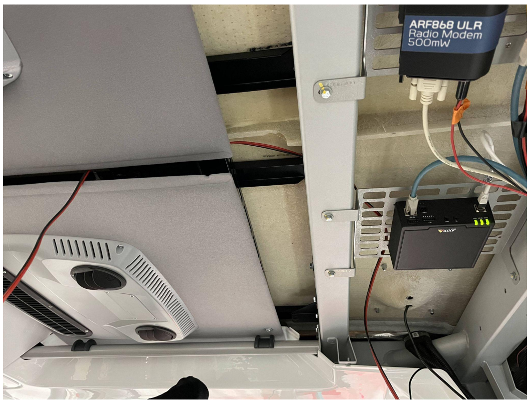
		 
11	Go to battery compartment and remove metel cover, to make room for the new power cables (being connected to the inverter)	

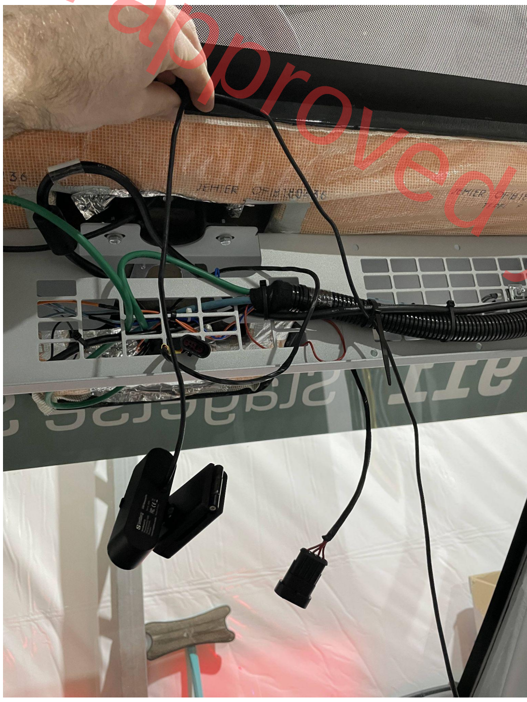
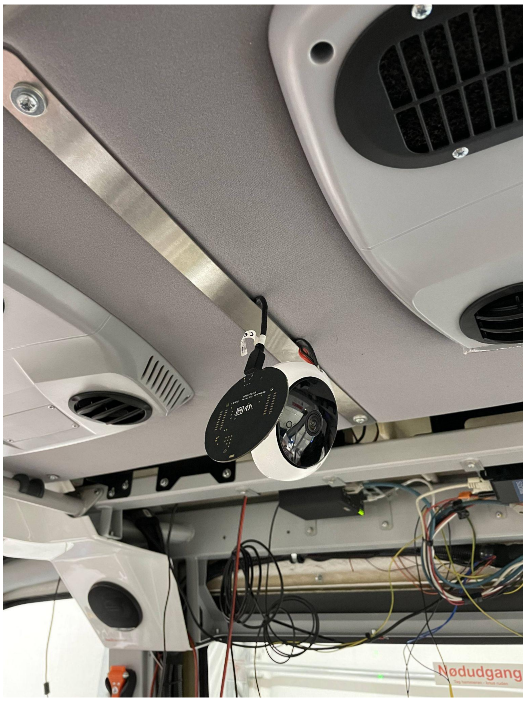
<table>
<tr><td></td><td></td><td></td></tr>
<tr><td>12</td><td>Place the inverter in the battery compartment (optional example with white box - not standard). BUT important to make sure the inverter is properly mounted and not able to move around in the compartment.</td><td></td></tr>
<tr><td>13</td><td>Wire the power cables from the battery compartment through the cord opening down into the wheelarch</td><td>
</td></tr>
</table>

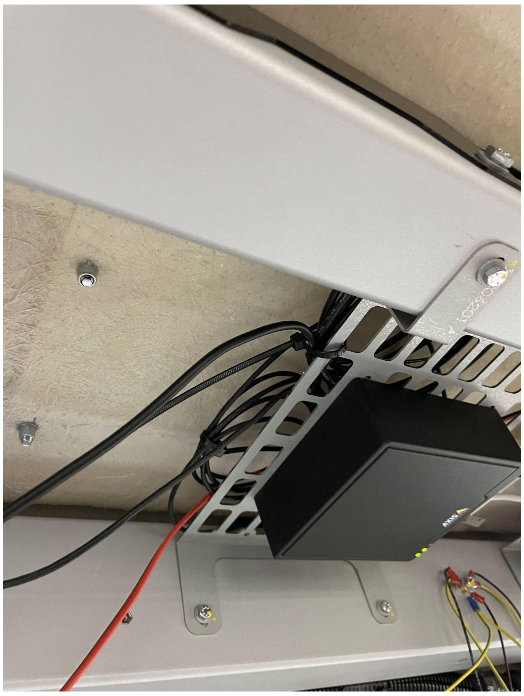
		
14	Wire the power cables back up into the shutte (try to get one through first - then tape the other one to the cord and pull both inside the vehicle). Use strips to hold the cords in the wheelarch area , as lose cables near the wheels are a bad idea.	 

		
15	Once the cables are inside the shuttle, getting in from behind the backrest of foldable seats. Use strips to fasten the power cables to the existing cables. Make sure the cables to not detach when driving to fasten good. Wire the power cables up vertically towards the ceiling. The jetson power cable al the way to the roof. The monitor cable just above the beam to the window.	
16	Once the cabels are fully wired through the wheelarch area. Close the cord opening (metal cover) and tidy up the battery compartment. Reverse the process of putting the wheel back on (down with lift, on with cover etc.)	No picture

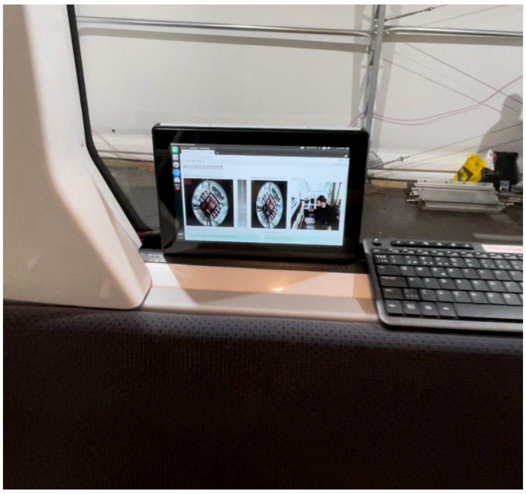
17	Place the jetson on the small deck to the left under the roof (use adhesive tape to mount it) and wire the jetson power cable up under the roof along side the beam, including the power cable for top camera (merged with the monitor power cable).	
18	Wire the HDMI cable from the jetson to the monitor down the beam. Use strips to contain the cable.	

19	Mount the monitor on the window with adhesive tape (M1 is the strongest adhesive tape). Connect the power cable.	
20	Wire the cables for the microphone and top camera from the jetson. Get above al lthe metal fittings under the rood and use strips to contain the cables. Connect the microphone to the jetson via the USB extension cable.	 

21	Position the webcam and wire the cables alongside the existing cables. Use strips to fasten the wire. Leve enough wire to put back the panel and have the webcam come out on top. Take the wire behind the side beam, up to the jetson and connect via the usb extension cable.	 
22	Once the length of the cables for the microphone and the top camera has been adjusted, close the top panels with the metal plate. Leave enough wire to position and tape (adhesive tape) the microphone and top camera to the metal plate. Watch out for pressing the cables toi much with the metal plate. Tight but not to cut the wires.	

23	Make sure that all the sensors and cameras are correeclt connected to the jetson, including the monitor and keyboard.	
24	Make sure that all cables around the entire system is mounted correct and not in dager of something or hanging loose. Use strips where it makes sense.	

25	Connecet the inverter to the battery and start up the jetson - check the entire system for connections and stability.	
26	Once everything works and has been tested. Screw back all the panels in the same orders as removed and you should now have completed the installation of the system. Final setup should look like this:	

10 API integration

10.1 API description and development

In this section, the API of the services is described along with the integration of the dashboard.

10.1.1 Dashboard integration

The services are launched via a unified launcher designed for Linux-based operating systems (Figure 85). The starting sequence involves the following steps:

- The launcher parses the configuration file ./config.py. Elements in the *shared[]* array are necessary modules required by the services. Elements in *modules[]* array are the service modules.
- The launcher imports the shared dependencies and starts them in the same memory page.
- Each service module is spawned from the main() function as a new process using *os.fork()*. Logic outside the *main()* entry point is never executed except when called.

Inter-process communication between the shared and service modules can be achieved using *av_ipc*. To send/receive IPC data, a PriorityQueue is defined in the main() entry point.

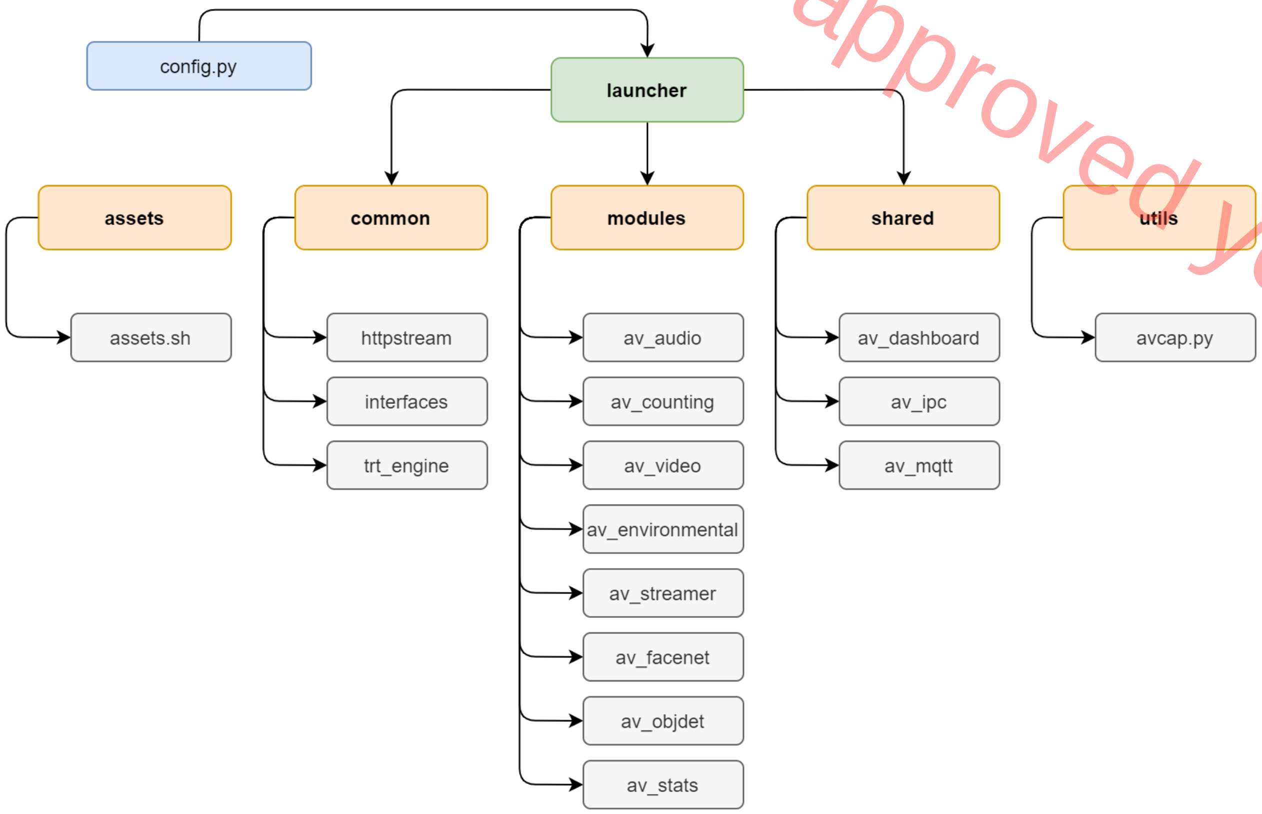


Figure 86. Unified services launcher for the AVENUE services

10.1.2 Third-party data transmission

The output of the services are transmitted in real-time using the *av_mqtt* shared module. The module connects to the MQTT broker provided by CERTH, performs subscription to the services topic and transmits the data that it receives from the other components. All of the services inherit the vehicle, service, data primary keys where vehicle contains information about the specific vehicle and service contains details about the specific service. A relevant output of the video and environmental modality is presented in Table 134 and Table 15 respectively along with their key description.

Table 14: JSON output of the *av_video* service.


```
{
    "vehicle":{
        "vehicle_id":{
            "model":"test",
            "vehicle_id":"test-vehicle-1",
            "location":"Thermi, Greece"
        }
    },
    "service":{
        "service_id":"av-video",
        "name":"Abnormal Event Detection Video"
    },
    "data":{
        "event_id":"3"
    }
```

}	
event_id	**Description**
"0"	Bagsnatch
"1"	Falldown
"2"	Fighting
"3"	Normal (no abnormal events)
"4"	Vandalism

Table 15: JSON output of the *av_audio* service.


```
{
    "vehicle":{
        "vehicle_id":{
            "model":"test",
            "vehicle_id":"test-vehicle-1",
            "location":"Thermi, Greece"
        }
    },
    "service":{
        "service_id":"av-audio",
        "name":"Abnormal Event Detection Audio"
    },
    "data":{
        "event_id":"0"
    }
}
```


Key	Type	Description
"0"	int	Background noise
"1"	int	Fighting
"2"	int	Scream

Table 16: JSON output of the *av_counting* service.


```
{
    "vehicle":{
        "vehicle_id":{
            "model":"test",
            "vehicle_id":"test-vehicle-1",
            "location":"Thermi, Greece"
        }
```

```
    },
    "service":{
        "service_id":"av-counting",
        "name":"Automated Passenger Presence"
    },
    "data":{
        "passengers":0,
        "safe_distance":true
    }
}
```

Key	Type	Description
"passengers"	int	Number of passengers detected in the shuttle
"safe_distance"	bool	True if the proximity of the passengers is safe. False otherwise.

Table 17: JSON output of the *av_facenet* service.


```
{
    "vehicle":{
        "vehicle_id":{
            "model":"test",
            "vehicle_id":"test-vehicle-1",
            "location":"Thermi, Greece"
        }
    },
    "service":{
        "service_id":"av-facenet",
        "name":"Follow my kid"
    },
    "data":{
        "passenger_list":[
            "Dimitris Tsiktsiris",
            "Anastasios Vafeiadis"
        ]
    }
}
```


The *passenger_list* key contains the names of the passengers detected as they appear on the service's database.

Table 18: JSON output of the *av_environmental* service.


```
{
    "vehicle":{
        "vehicle_id":{
            "model":"test",
            "vehicle_id":"test-vehicle-1",
            "location":"Thermi, Greece"
```

```
        }
    },
    "service":{
        "service_id":"av-counting",
        "name":"Automated Passenger Presence"
    },
    "data":{
        "smoking":"No",
        "temp_forecast":null,
        "aq_index":95,
        "temp":2.15,
        "humid":49.35,
        "co2":504,
        "voc":70,
        "pm25":15
    }
}
```


Key	Type	Description
"smoking"	string	"Yes" if smoke is detected - "No" otherwise
"temp_forecast"	float	Predicted Celsius degrees for the next 5 minutes
"aq_index"	int	Air quality index (0-100). 100 indicates the best air quality
"temp"	float	Current temperature in Celsius degrees
"humid"	float	Current humidity percent
"co2"	int	Current Carbon Dioxide / 400-5,000 ppm
"voc"	int	Current Volatile Organic Compounds (VOCS) / 0-60 ppm
"pm25"	int	Current Particulates (PM2.5) / 0-1,000 µg/m³

Table 19: JSON output of the *av_handgesture* service.


```
{
    "vehicle":{
        "vehicle_id":{
            "model":"test",
            "vehicle_id":"test-vehicle-1",
            "location":"Thermi, Greece"
        }
    },
    "service":{
        "service_id":"av-handgesture",
        "name":"Hand gesture module"
    },
```

```
    "data":{
        "gesture":-1
    }
}
```

Key	Type	Description
“gesture”	int	-1 no gestures detected 0 thumbs down detected 1 thumbs up detected

11 Validation of sensor-deployed solutions

This Section presents the validation of the in-vehicle services proposed for the AVENUE project. It includes the validation results of the visual, acoustic and environmental modalities as well as future recommendations for the deployed services.

All the services were validated using accuracy and F1-Score. The accuracy is calculated as the fraction of the total correct samples over the total samples, while the F1-Score provides a more representative metric in the case of unbalanced data.

11.1 Enhance the sense of security and trust

In this section the validation results and future recommendations for the service are presented.

11.1.1 Validation procedure

The results were validated during a two-fold live session on Monday Aug 29, 2022 and Wednesday Aug 31, 2022. During the live meeting, staff from HOLO performed fighting, bag snatch, falling down and vandalism scenarios in vehicle P109 that serves the route Slagelse in Denmark. CERTH and MobileThinking participated in the validation meeting by monitoring the services and operator dashboards correspondingly.

11.1.2 Validation result

During the live sessions, 20 video segments of 3 seconds with abnormal events and 20 segments with normal data were inferenced in real-time. Specifically, in Table 20, the number of samples validated for each class are shown and in Table 21, the accuracy and the F1-Score metrics are calculated based on the inference result.

Table 20: Number of samples validated for each class

Event	Samples
Bagsnatch	5
Fighting	4
Falldown	6
Vandalism	5

Table 21: Accuracy and F1-Score for each class

Event	Accuracy	F1-Score
Bagsnatch	0.9	0.9091
Fighting	0.8889	0.8889
Falldown	0.8333	0.8571

Vandalism	0.9167	0.9091

The results indicated that the algorithms were able to correctly identify most of the performed scenarios with 89% accuracy and the appropriate notifications were also captured in the operator's dashboard. The two dashboards were correctly synchronised regarding the real-time event detection. Figures 87 and 88 show the two dashboards from CERTH and MobileThinking respectively regarding the results for the service "Enhance the sense of security and trust". Add inference results from videos

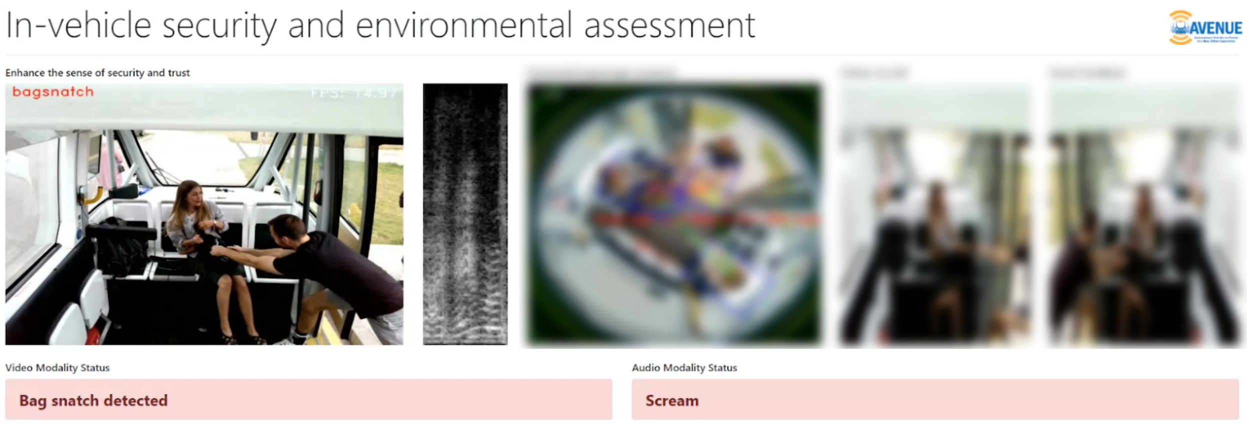


Figure 87: CERTH frontend interface regarding the service "Enhance the sense of security and trust".

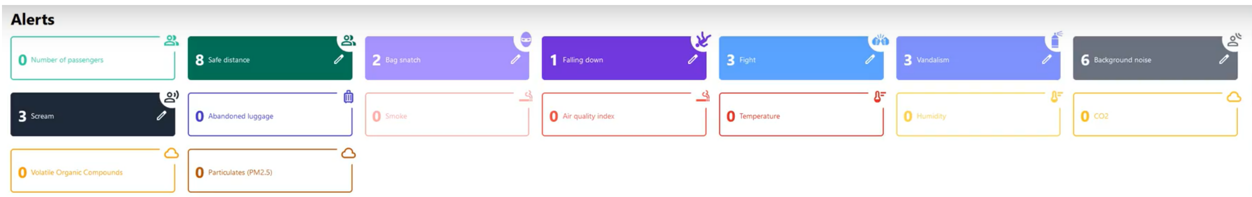


Figure 88: MobileThinking dashboard regarding the service "Enhance the sense of security and trust".

11.1.3 Recommendations

As the services were deployed in real pilot sites but in care units, the number of abnormal events (e.g. fighting) could only be simulated by the HOLO operators. The CERTH algorithms could be further enhanced in order to further reduce the false positives. This could be possible via additional real-world data capture with more abnormal events such as fighting and bag snatch. The developed algorithms are designed in order to be capable of easily adding new scenarios. For instance, if a drunk or a sick person gets onboard the autonomous shuttle, the algorithm could detect if that person vomited and trigger an alert.

11.2 Automated passenger presence

The in-shuttle operator on site has already manually been counting passengers using the operator app, hence validating the automated passenger counting has been done through comparison with data received from the OP phone's data stream. In Figure 89 below the manual passenger count is seen to the left, where each person getting on the shuttle is entered as 1s in the data stream. Within the same minute as the operator manually counts the 4 entries, the data stream received from the jetson increases to a count of 5 passengers (4 passengers and the operator).

Row	timestamp	qty
11	2022-07-26 12:20:09.609000 U...	1
12	2022-07-26 12:20:08.858000 U...	1
13	2022-07-26 12:20:08.487000 U...	1
14	2022-07-26 12:20:08.251000 U...	1

Row	timestamp	passengers...
68	2022-07-26 12:22:39.253144 U...	5
69	2022-07-26 12:21:39.427374 U...	5
70	2022-07-26 12:20:47.357969 U...	5
71	2022-07-26 12:20:15.349922 U...	4
72	2022-07-26 12:19:45.832723 U...	1
73	2022-07-26 12:18:58.261609 U...	1
74	2022-07-26 12:18:51.749958 U...	1
75	2022-07-26 12:18:19.611659 U...	1

Figure 89: Screenshots from BigQuery: 4 passengers are getting on the shuttle

Right: Manual count from OP-app: data points only received when state changes (button is pushed in the shuttle)
Left: Automated passenger presence from CERTH equipment: data points received continuously. The count includes the on-board operator.

In Figure 90 below the manual count is seen to decrease by individual -1 data points, when passengers leave the shuttle again. As for the automated count, the passenger number gradually returns back to 1. Comparing the two data streams, it is clear that the continuous automated count uses significantly more storage space compared to the manual count, which only adds up to a hundred bytes a day, whereas a single day of the automated data is easily several megabytes of data.

Row	timestamp	qty
7	2022-07-26 12:32:34.852000 U...	-1
8	2022-07-26 12:32:34.668000 U...	-1
9	2022-07-26 12:32:34.508000 U...	-1
10	2022-07-26 12:32:34.300000 U...	-1
11	2022-07-26 12:20:09.609000 U...	1
12	2022-07-26 12:20:08.858000 U...	1
13	2022-07-26 12:20:08.487000 U...	1
14	2022-07-26 12:20:08.251000 U...	1

45	2022-07-26 12:35:08.134689 U...	1
46	2022-07-26 12:35:01.472960 U...	1
47	2022-07-26 12:34:30.941244 U...	1
48	2022-07-26 12:34:16.242700 U...	1
49	2022-07-26 12:33:15.324778 U...	1
50	2022-07-26 12:32:30.873051 U...	3
51	2022-07-26 12:32:22.474976 U...	4
52	2022-07-26 12:31:48.796350 U...	5

Figure 90: Screenshots from BigQuery: 4 passengers are getting off the shuttle

However from the manual comparison of the 2 distinct events in the data streams above, it becomes clear that the automated count seems precise in terms of timestamp and accurate in terms of the passenger count.

11.2.1 Validation result

Comparing all data points received within a stable 2 day period of operation, the accuracy of the count was investigated further. Only the times where data from the CERTH equipment showed more than 1 person in the shuttle was extracted and compared. Hence, be aware of the jumps in the timestamps on the x-axis. The timestamp is given in UTC time, meaning the actual time was (Copenhagen summer time) is 2 hours ahead.

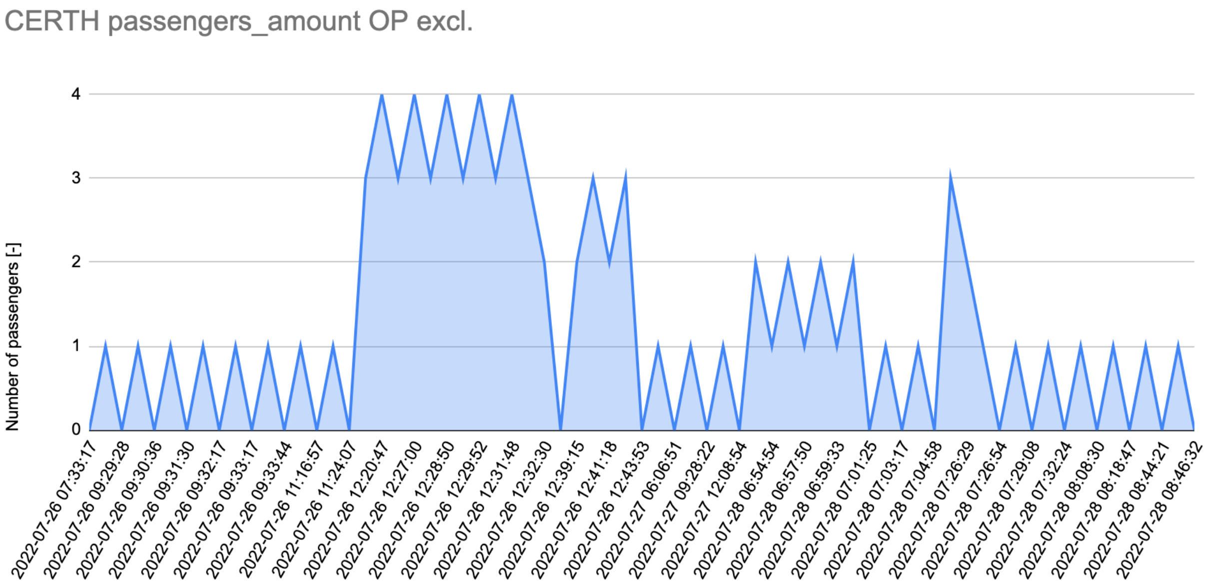


Figure 91: Passenger count over time received from the CERTH data stream

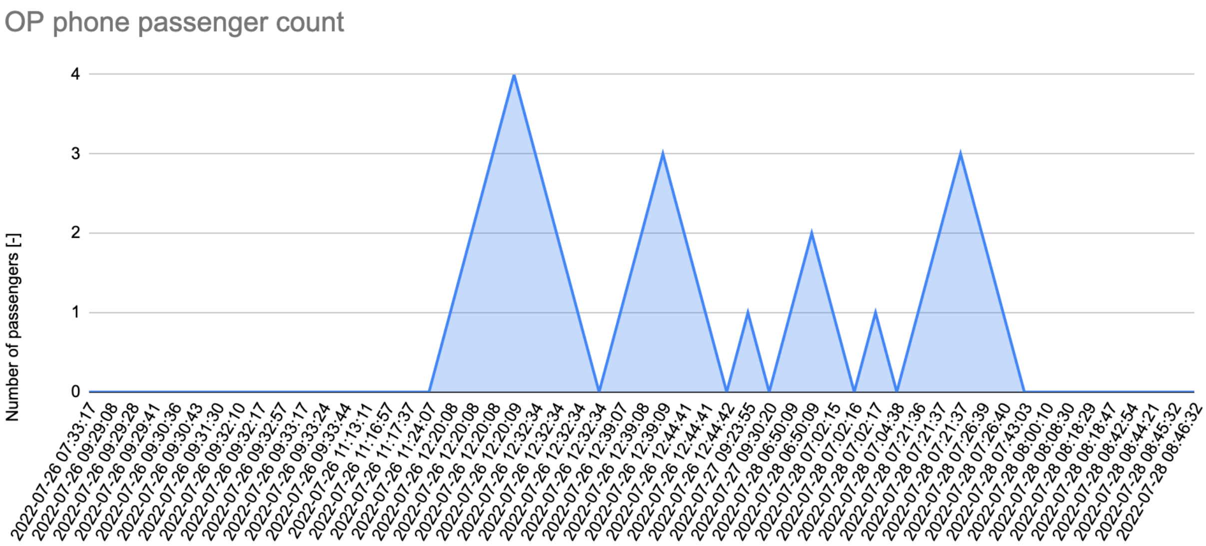


Figure 92: Passenger count over time received from manual count from the operator phone

Visually comparing the two data streams, it is seen that the patterns show a similar count of passengers. Initially the algorithm below was developed to determine the total number of passengers on board, counting only the increases in number of passengers.

$$\begin{gathered} if\ r_i > r_{i-1} \\ then\ n_i = n_{i-1} + r_i - r_{i-1} \\ else\ n_i = n_{i-1} \end{gathered}$$

Where

r_i is the number of passengers present in time i and

n_i is the total number of passengers at time i

However, the oscillating nature of the automated counting data will result in a higher count of total passengers, as extra passengers are added to the total every time an oscillation occurs.
Moreover the exact number of passengers in the shuttle, at what times, becomes inaccurate, hence for more analytical purposes such a data stream is very challenging to work with.

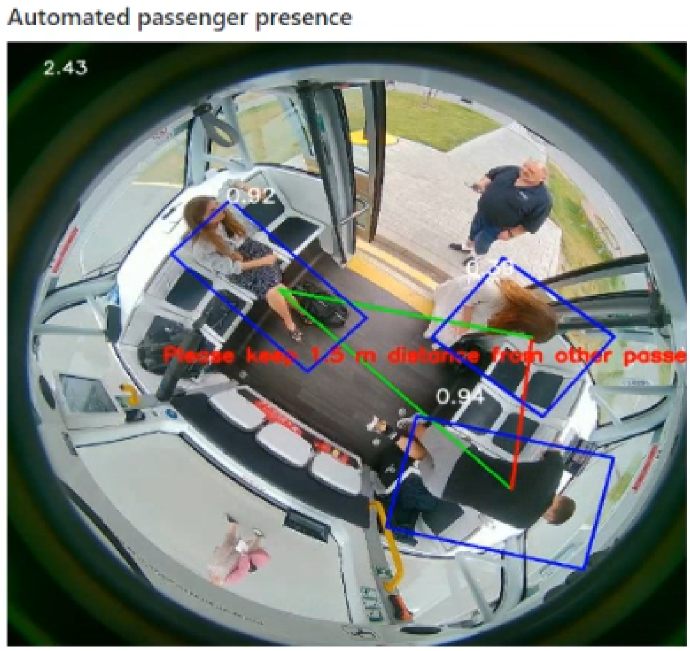


Figure 93. CERTH feed regarding the service “Automated passenger presence”.

11.2.2 Recommendations

For future setup of data streams to use for automated passenger counting some verification and cross-validation of the data would be of great value, as the data in a stand-alone format is too fragile. As an example, cross validation with data points received on the status of the shuttle door from the vehicle API or from the sensors. This could be used to “lock” the passenger count, when doors are closed, as passengers in reality are not able to enter or exit the shuttle. Moverover, it could be an option to filter out oscillations in the count within short time periods or by using the most present count between to timestamps with oscillations between them. However further investigation on what time interval would be appropriate, as passengers can enter and exit the shuttle for short periods of time, without actually driving with the shuttle and vice versa.
Additionally, the many data points received from the CERTH equipment compared to manually entered data, makes data processing rigid and consuming.

11.3 Follow my kid service

In this section the validation results and future recommendations for the service are presented.

11.3.1 Validation procedure

The results were validated during a two-fold live session on Monday Aug 29, 2022 and Wednesday Aug 31, 2022. CERTH and MobileThinking participated in the validation meeting by monitoring the services and operator dashboards correspondingly.

11.3.2 Validation results

During the live sessions, 25 video segments of 3 seconds with facial data of 5 passengers enrolled into the service and 25 segments with unknown passengers were inferenced in real-time. Specifically, in Table 22, the number of samples validated for each class are shown and in Table 23, the accuracy and the F1-Score metrics are calculated based on the inference result.

Table 22. Number of samples validated for each class

Gender	Samples
Enrolled male	3
Enrolled female	2
Unknown male	5
Unknown female	4

Table 23. Accuracy and F1-Score for each class

Event	Accuracy	F1-Score
Enrolled male	0.8571	0.8571
Enrolled female	0.8	0.8
Unknown male	0.9167	0.9091
Unknown female	0.875	0.8889

The results indicated that the algorithms were able to correctly identify people from HOLO that were enrolled in the service with 86% accuracy. Figure 93 shows the dashboard from CERTH regarding the results for the service “Follow my kid”.

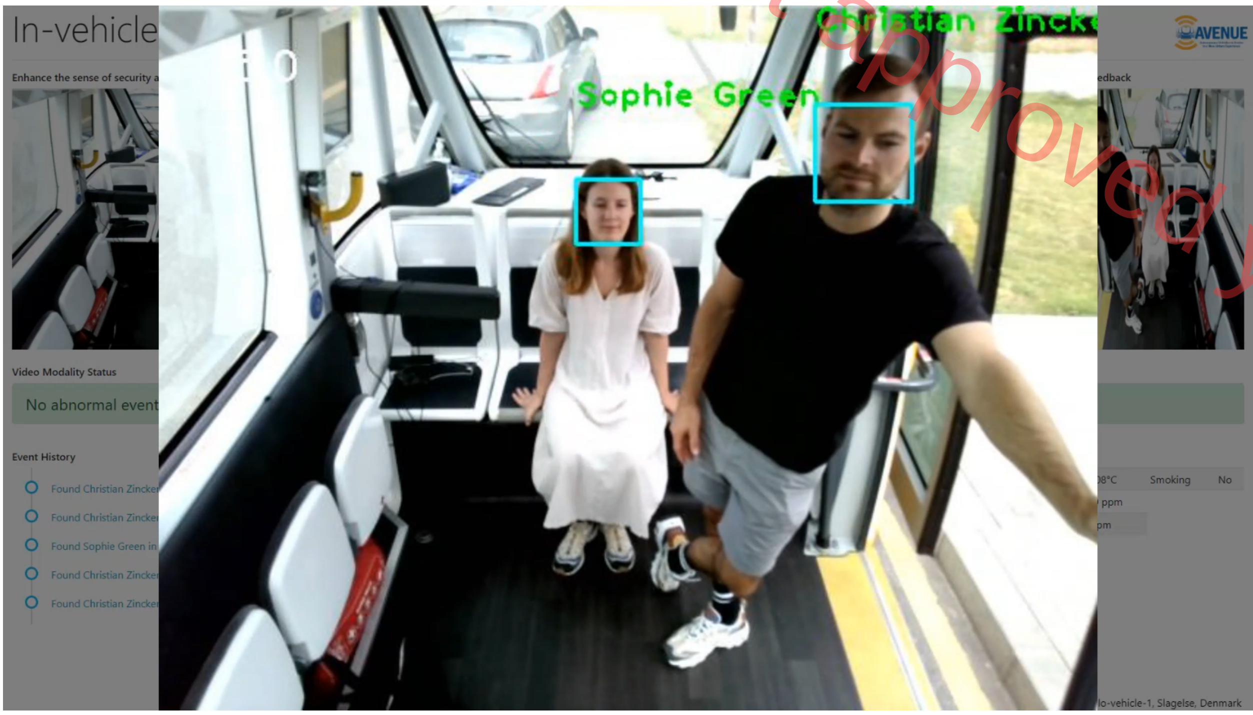


Figure 94. CERTH frontend interface regarding the service “Follow my kid”.

11.3.3 Recommendations

A supplementary smartphone app could be integrated into the existing solution in order to provide the user the ability to use photographs for facial verification.

11.4 Shuttle environment assessment

In this section the validation results and future recommendations for the service are presented.

11.4.1 Validation procedure

The results were validated during a two-fold live session on Monday Aug 29, 2022 and Wednesday Aug 31, 2022. CERTH and MobileThinking participated in the validation meeting by monitoring the services and operator dashboards correspondingly.

11.4.2 Validation result

During the live sessions, two cigarettes were lit inside the shuttle. The service detected the increasing number of VOCs and PM2.5 particles and triggered a smoking alert. The event was also sent to the operator's dashboard successfully.

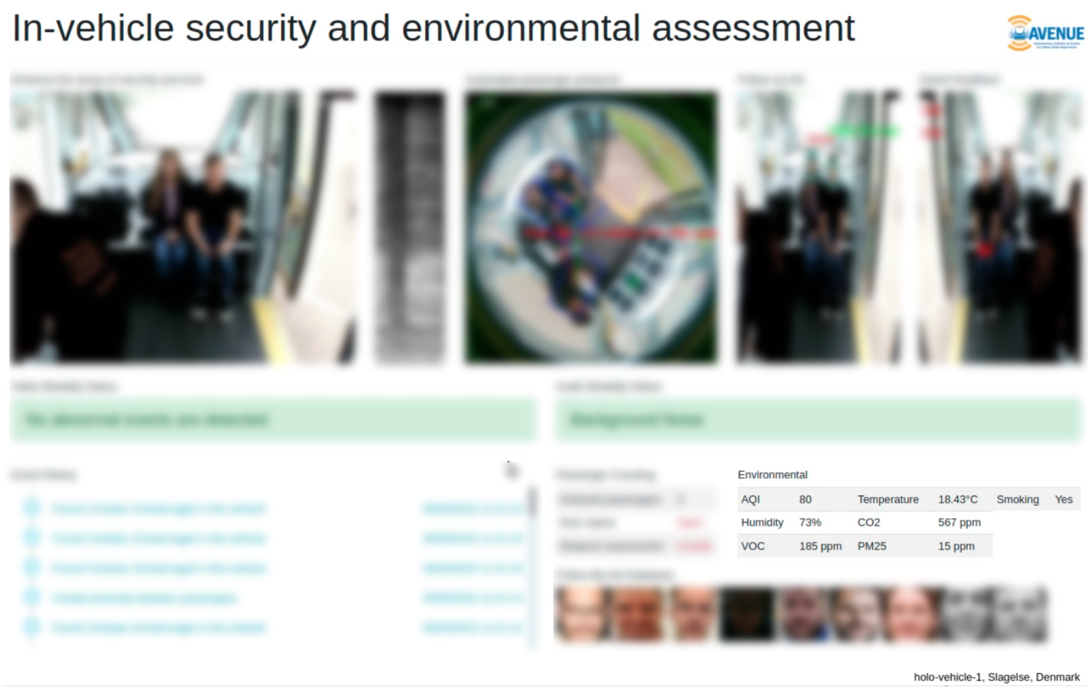


Figure 95. CERTH frontend interface regarding the service "Shuttle environment assessment".

Figure 96. CERTH frontend interface regarding the service "Shuttle environment assessment".

11.4.3 Recommendations

In the future, the Awair sensor could be replaced with a custom hardware solution which is more compact. The solution will be able to provide the same metrics and would be more seamlessly integrated into the autonomous vehicle.

11.5 Smart feedback system

In this section the validation results and future recommendations for the service are presented.

11.5.1 Validation procedure

The results were validated during a two-fold live session on Monday Aug 29, 2022 and Wednesday Aug 31, 2022. During the live meeting, staff from HOLO performed thumbs up and thumbs down scenarios in vehicle P109 that serves the route Slagelse in Denmark. CERTH and MobileThinking participated in the validation meeting by monitoring the services and operator dashboards correspondingly.

11.5.2 Validation results

During the live sessions, 16 video segments of 3 seconds with hand gestures and 16 segments with no gesture were inferenced in real-time. Specifically, in Table 24, the number of samples validated for each class are shown and in Table 25, the accuracy and the F1-Score metrics are calculated based on the inference result.

Table 24. Number of samples validated for each class

Event	Samples
Thumbs up	10
Thumbs down	6

Table 25. Accuracy and F1-Score for each class

Event	Accuracy	F1-Score
Thumbs up	0.95	0.9524
Thumbs down	0.9286	0.9231

The results indicated that the algorithms were able to correctly identify most of the aforementioned gestures with 94%. Figure 96 shows the dashboard from CERTH regarding the results for the service “Smart feedback system”.

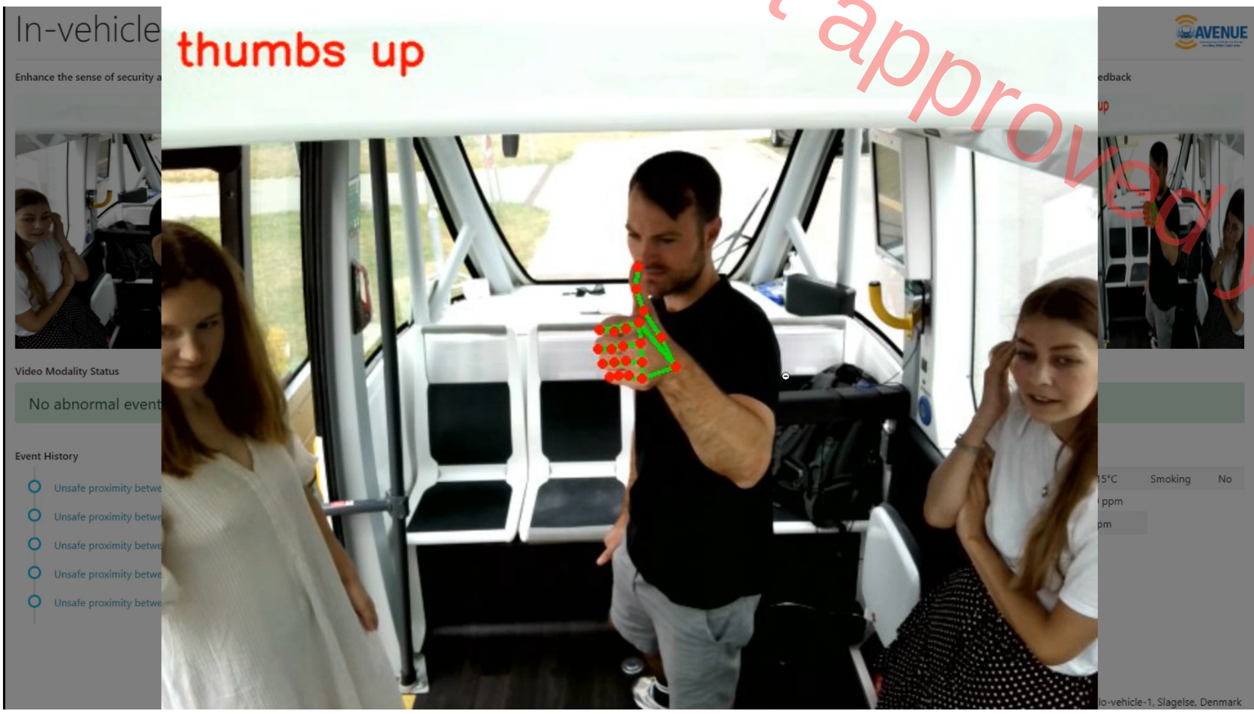


Figure 97. CERTH frontend interface regarding the service "Smart feedback system".

11.5.3 Recommendations

In the future the algorithms could be improved with a rule based method in order to exclude multiple gestures from the same person. With additional data, the algorithm could be able to detect additional gestures (e.g., stop) that in conjunction with other parameters or sensor data (video or audio feed) could help to enhance the passenger experience and safety.

12 Conclusions

Five in-vehicle services have been developed, prototyped and tested in the AVENUE project over many phases in collaboration between CERTH, Amoblity and MobileThinking.

Each service targets different essential functions currently present in the vehicles and performed by the safety steward. As an important objective of AVENUE is to prove what it takes to remove the safety steward an early analysis was made to understand the role of the safety steward. The services target safety and trust inside the vehicles looking towards a future without the safety steward. As the industry aims to find ways to remove the safety stewards the 5 services have been developed and tested as an initial way to reduce the role of the safety steward and prepare for fully autonomous driving in the future.

All the services are centred around camera and sensor systems inside the shuttle. The prototyping and testing of the 5 services have over the last 12 months of the project been matured to a level where the project can offer a close-to-ready market solution for AVs with a relatively low cost and implementation effort. The services have been stabilised and made robust for the sake of the future, with both the quality of the services in mind but also with cyber security in mind. Based on the large efforts of the partners involved in the development and testing some very direct recommendations and objectives have been generated - ensuring the best potential use and further development of the services for the future.

It is expected that the solutions will mature as the autonomous technology will represent a large part of the transport means in public as well as private transport systems.

13 Recommendations

As an outcome of the development, testing and validation process, the project has come to the following recommendations regarding the deployment and implementation of in-vehicle services. As described in the conclusion the potential value of the services is perceived to be very high and the project aims to inspire new projects to follow up on the learnings and following recommendations for the future.

Enhance the sense of security and trust:
As the services were deployed in real pilot sites but in care units, the number of abnormal events (e.g. fighting) could only be simulated by the HOLO operators. The CERTH algorithms could be further enhanced in order to further reduce the false positives. This could be possible via additional real-world data capture with more abnormal events such as fighting and bag snatch. The developed algorithms are designed in order to be capable of easily adding new scenarios. For instance, if a drunk or a sick person gets onboard the autonomous shuttle, the algorithm could detect if that person vomited and trigger an alert.

Automated passenger presence:
For future setup of data streams to use for automated passenger counting some verification and cross-validation of the data would be of great value, as the data in a stand-alone format is too fragile. As an example, cross validation with data points received on the status of the shuttle door from the vehicle API or from the sensors. This could be used to “lock” the passenger count, when doors are closed, as passengers in reality are not able to enter or exit the shuttle. Moverover, it could be an option to filter out oscillations in the count within short time periods or by using the most present count between to timestamps with oscillations between them. However further investigation on what time interval would be appropriate, as passengers can enter and exit the shuttle for short periods of time, without actually driving with the shuttle and vice versa.
Additionally, the many data points received from the CERTH equipment compared to manually entered data, makes data processing rigid and consuming.

Follow my kid service:
A supplementary smartphone app could be integrated into the existing solution in order to provide the user the ability to use photographs for facial verification.

Shuttle environment assessment:
In the future, the Awair sensor could be replaced with a custom hardware solution which is more compact. The solution will be able to provide the same metrics and would be more seamlessly integrated into the autonomous vehicle.

Smart feedback system:
In the future the algorithms could be improved with a rule based method in order to exclude multiple gestures from the same person. With additional data, the algorithm could be able to detect additional gestures (e.g., stop) that in conjunction with other parameters or sensor data (video or audio feed) could help to enhance the passenger experience and safety.