European Commission

Horizon 2020
European Union funding
for Research & Innovation

Big Data technologies and extreme-scale analytics

Multimodal Extreme Scale Data Analytics for Smart Cities Environments

D2.1: Collection and Analysis of Experimental Data[†]

Abstract: Project MARVEL will create and publicly share with the academics, industrial community, and smart cities, a data pool of experimental multimodal audio-visual data and will showcase the use of the data and its processing across various pilots. This report documents the process for the collection and analysis of the experimental data. The use cases and AI tasks required to process the audio-visual data and enable the implementation of the pilots are first described. This knowledge determines how and where the audio-visual data is collected and annotated. The various devices, namely microphones and cameras, and their deployment are described next, followed by software tools that will be used in the data annotation task, that will be carried out according to what is required for training the AI models. The data is analysed to determine which parts constitute personal data, followed by a discussion on the appropriate data anonymisation techniques, which should ensure the sharing of GDPR compliant data. In addition, the data value chains are defined, including the data owner and access rights at each processing stage. The proposed datasets and AI models are matched and compared to the use cases and any gaps are identified. The volume and velocity at which data is collected and moved from one network layer to another are estimated from the technical specifications of the devices as well as from the expected output of the processing stages. These estimates enable the initial planning of the MARVEL framework, which promises a solution to collect and process big data of high volume and high variety while optimising both flow and processing at any appropriate point of the edge-fog-cloud infrastructure, typical of a smart city.

Contractual Date of Delivery	30/06/2021
Actual Date of Delivery	30/06/2021
Deliverable Security Class	Public
Editor	*Adrian Muscat (GRN)*
Contributors	IFAG, AU, ATOS, CNR, INTRA, FBK, AUD, TAU, MT, UNS, ITML, GRN
Quality Assurance	*Manolis Marazakis (FORTH)* *Grigorios Kalogiannis (STS)*

[†] *The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957337.*

The *MARVEL* Consortium

Part. No.	Participant organisation name	Participant Short Name	Role	Country
1	FOUNDATION FOR RESEARCH AND TECHNOLOGY HELLAS	FORTH	Coordinator	EL
2	INFINEON TECHNOLOGIES AG	IFAG	Principal Contractor	DE
3	AARHUS UNIVERSITET	AU	Principal Contractor	DK
4	ATOS SPAIN SA	ATOS	Principal Contractor	ES
5	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR	Principal Contractor	IT
6	INTRASOFT INTERNATIONAL S.A.	INTRA	Principal Contractor	LU
7	FONDAZIONE BRUNO KESSLER	FBK	Principal Contractor	IT
8	AUDEERING GMBH	AUD	Principal Contractor	DE
9	TAMPERE UNIVERSITY	TAU	Principal Contractor	FI
10	PRIVANOVA SAS	PN	Principal Contractor	FR
11	SPHYNX TECHNOLOGY SOLUTIONS AG	STS	Principal Contractor	CH
12	COMUNE DI TRENTO	MT	Principal Contractor	IT
13	UNIVERZITET U NOVOM SADU FAKULTET TEHNICKIH NAUKA	UNS	Principal Contractor	RS
14	INFORMATION TECHNOLOGY FOR MARKET LEADERSHIP	ITML	Principal Contractor	EL
15	GREENROADS LIMITED	GRN	Principal Contractor	MT
16	ZELUS IKE	ZELUS	Principal Contractor	EL
17	INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK	PSNC	Principal Contractor	PL

Document Revisions & Quality Assurance

Internal Reviewers

1. *Manolis Marazakis (FORTH)*
2. *Grigorios Kalogiannis (STS)*

Revisions

Version	Date	By	Overview
1.0	30/06/2021	Editor, IR, SPTM, PC	Review and approval
0.4.3	30/06/2021	Editor	Addressed 4th round of reviews (STPM)
0.4.2	29/06/2021	Editor	Addressed 3rd round of reviews (PC)
0.4.1	24/06/2021	Editor	Addressed 2nd round of reviews (FORTH, STS)
0.4.0	21/06/2021	Editor	Addressed reviewers' comments (FORTH, STS)
0.3.1	9/06/2021	Editor	Final Draft
0.3.0	01/06/2021	Editor	First draft
0.2.2	09/05/2021	Editor	Final ToC
0.2.1	01/05/2021	WPL, STPM	Comments on the ToC
0.1.0	21/04/2021	Editor, (GRN)	ToC

Disclaimer

The work described in this document has been conducted within the MARVEL project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957337. This document does not reflect the opinion of the European Union, and the European Union is not responsible for any use that might be made of the information contained therein.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AAC	Automated Acoustic Capturing
AI	Artificial Intelligence
ASC	Acoustic Scene Classification
AV	Audio-Visual
BB	Bounding Box
CSV	Comma Separated Variables
dBSPL	decibel Sound Pressure Level
DL	Deep Learning
DMP	Data Management Plan
DMT	Data Management Toolkit
EC	European Commission
E2F2C	Edge-to-Fog-to-Cloud
FAIR	Findable, Accessible, Interoperable, and Re-usable
FL	Federated Learning
FPS	Frames Per Second
GA	Grant Agreement
GB	Gigabyte
GDPR	General Data Protection Regulation
HPC	High Performance Computing
IoT	Internet of Things
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
MAVD	Monte-Video Audio-Video Data
MB	Megabyte
MKV	Matroska Video File Format
MP3	Moving Pictures Video File Format
ML	Machine Learning
PbD	Privacy by Design
PD	Personal data
SED	Sound Event Detection
SELD	Sound Event Localisation and Detection
SOTA	State-of-the-art
VAD	Voice Activity Detection
WAV	Waveform Audio File Format
WP	Work Package
XML	Extensible Markup Language
YOLO	You-Only-Look-Once

Executive Summary

Advances in digital technologies are fast becoming the key elements in the creation and augmentation of Big Data. Smart cities are one of the primary data pools, where data is generated from a very large number of IoT devices and sensors, that are diverse not only in terms of volume but also in terms of variety, velocity, veracity, and value. To this present situation, one of the biggest challenges is to extract commercial and meaningful knowledge from this data. This presents an opportunity for developing new methodologies, techniques, and tools for information extraction and manipulation that are different from the traditional ones. Towards these aspects, project MARVEL aims, on one hand, to harmonise techniques and technologies in the areas of Artificial Intelligence (AI), analytics, multimodal perception, software engineering, High Performance Computing (HPC), and typical architectural approaches in processing heterogeneous and distributed data in smart cities environments. Furthermore, and on the other hand, project MARVEL encourages the vision of EU Data Economy not only by augmenting and sharing a Data Corpus with the international scientific and research community driving innovation in multimodal processing and analytics but also solving issues that may arise in the Big Data Value chain in terms of acquisition, storage, analysis, and visualisation.

Project MARVEL deals with the implementation of a data management toolkit (DMT) that manages the flow and processing of data over a complex Edge-to-Fog-to-Cloud (E2F2C) infrastructure. The platform receives data from a variety of IoT devices and allows the consumer (researchers or engineers developing an application) to process the data using built-in AI models. The platform processes the data either centrally or in a distributed manner. The platform will be showcased over a number of use cases executed in the municipalities of Trento, Malta, and Novi Sad. This requires the collection of datasets and the development of relevant AI tasks. In addition, MARVEL will publicly share with the academics, industrial community, and smart cities, a data pool of experimental multimodal audio-visual data. This report documents the process for the collection and analysis of the experimental data. This document focuses on two types of datasets. Those, that are often annotated with class labels, etc, used during the training of the models, and datasets that are either the raw data itself or the output of the models during inference.

The definition and elaboration of the use cases determine the type of data and AI tasks that are necessary to implement the pilots. Use cases often share some of the data and AI tasks and this characteristic allows for the optimisation of resources and respective processing pipelines. It is therefore important to have an overall view of the data and the AI tasks and how these interact to implement the pilots. This information helps the consumer of the data (e.g., the pilots in project MARVEL) in the selection of the appropriate methods to process the data with.

The audio-visual data are collected from the IoT devices (namely microphones and cameras) and the velocity and volume at which data is delivered impacts the way the data is processed over the E2F2C infrastructure. The type of devices and their characteristics and how data is processed (AI tasks) determine the data value chain, i.e., how the data is transformed at different stages of processing, most importantly the type and nature of the data and the volume per unit time. All this information is necessary when the platform is distributing data and processes to deliver the smart city services.

Most of the data that is used during the training of AI models that take as input sound, video or both is annotated and curated. This task requires the use of annotation tools and the collection of data, whether from staged (the use of actors) experiments or recordings of real-life. Annotation tools that are readily available and potentially suitable for labelling the training

datasets that will be used to train the AI models in MARVEL are described. It is also important to include metadata that records information on the technical characteristics of the devices, location and the date and time of data collection, data format, and labelling schema. The choice of appropriate keywords should ensure that the datasets adopt the FAIR (Findable, Accessible, Interoperable, and Re-usable) principles. Moreover, project MARVEL will publicly share datasets and therefore information on who the data owner is, who has access to the data, terms of use, whether ownership changes as the data is processed and other aspects of legal and commercial nature are defined.

The edge devices (microphones and cameras) will be recording real-life and invariably will also record Personal Data (PD), which is not necessary for the execution of the pilots. The PD present in the dataset are identified and AI models to anonymise the audio and video recordings are suggested to guide the development of such models in the respective work packages (WPs). The anonymisation model will ensure privacy and adherence to General Data Protection Regulation (GDPR) legislation. For the case of the data collection process during staged experiments, all participants in the experiment will sign a written consent form and therefore anonymisation techniques may not be needed for these specific cases. In addition, it is the intention of the MARVEL project to carry out anonymisation at the edge, thus simplifying system complexity and reducing the risk of personal data breach.

Finally, on the basis of the available information, it is possible to start planning on how the MARVEL framework (data management platform) will deliver a solution that handles both the flow and the processing of data of high volume and variety over the E2F2C infrastructure and identify gaps in technology that need to be addressed in; WP2, T2.2-Data Management and Distribution, and T2.4 – Sharing multimodal Corpus-as-a-Service; WP3, T3.4-Adaptive E2F2C distribution and optimisation of AI tasks; and WP5–Infrastructure Management and Integration.

1 Introduction

The smart city concept has been developed to tackle the challenges society is currently facing and will likely face in the not-so-distant future. Most of these problems or challenges are a by-product of economic growth and progress. Smart cities depend on the development of systems that rely on the delivery of timely and detailed data, to deliver services that benefit the citizen and improve the overall quality of life. Examples of such applications can be found in the areas of healthcare, transport, and law enforcement. At the heart of the smart city are efficient methods of collecting, processing, storing, and transferring data to implement new applications. One of the objectives of the MARVEL project is to create and publicly share processed multimodal audio-visual data to stimulate research in audio-visual processing and drive the industrial development of new applications for smart cities. Project MARVEL also implements example use cases to showcase its framework.

This document reports on the processes involved in the collection of the multi-modal data. Section 1 introduces the document and its relation to project MARVEL. Section 2 discusses potential use cases and the algorithms or AI tasks that are needed to implement the use cases, whilst section 3 describes the data collection process for collecting an initial set of experimental datasets and outlines the plan to incrementally augment the datasets. Section 4 describes the datasets in hand and section 5 analyses the datasets and discusses requirements for the data management framework. Section 6 discusses algorithmic methods that are useful in the anonymisation of the datasets and section 7 concludes the report.

1.1 Purpose and scope of the document

The main purpose of this document is to report on the process followed in the collection of MARVEL experimental distributed data assets (provided by GRN, MT, UNS), i.e., the type of data generated as a function of the devices deployed and the type of processing (including AI tasks) carried out on the data. Together these define how the data is transformed as it flows through the various processing/storage stages until it reaches its final destination and stored for further processing. Together, these sensory, processing, and storage components define the characteristics of the distributed data assets. This necessitates the definition of the data value chains, including ownership and any access control rights, accessibility and all relevant technical, organisational, legal and commercial aspects of data sharing to the consortium.

Following the definition of the available data value chains for the execution of the experiments, the datasets are analysed in depth to inform the necessary processing in the WPs that follow. Several aspects are considered such as the data model, format, velocity, volume, variety, and schema that data conforms to.

In addition, it is to be expected that the microphones and cameras deployed in smart cities environments and the mobile technologies distributed to several people participating in project's experiments will occasionally capture personal data (PD), which are of no use to the experiments. However, the data capturing process and algorithms must be compliant with the GDPR regulation and address all privacy concerns and for this reason, this document will guide the development of the respective tools and algorithms that deal with private data and anonymisation.

1.2 Contribution to WP2 and project objectives

The scope of WP2 (MARVEL multimodal data Corpus-as-a-Service for smart cities) is to fulfil one of the objectives of the MARVEL project. i.e., to create and publicly share with the academics, industrial community and smart cities, a data corpus of processed multimodal audio-

visual data, mainly to stimulate research on multimodal audio-visual analytics by overcoming the lack of publicly available data and to enable smart cities to build and deploy innovative applications that are based on multimodal perception and intelligence. WP2 will therefore (i) develop a Data Management and Distribution Toolkit, which is necessary for the handling of massive amounts of data coming from various sources and in dealing with their management and proper distribution (T2.2), thus facilitating public access to data (T2.4) and (ii) implement an incremental scheme for the continuous augmentation of the dataset (T2.3) with data that is GDPR compliant through the consideration of ethical and privacy concerns (T2.5). This document paves the way to the above two objectives.

1.3 Relation to other work packages and deliverables

This document is related to other work packages and tasks in the project as follows;

T1.3 in WP1 defines the experimental protocol, i.e., real-life societal trial cases in smart cities environments (Malta, Trento, Novi Sad) and elaborates on the use cases as well as the data collection equipment, which are both described in this document.

In WP3, T3.1 is concerned with the development of AI-based methods for data privacy; T3.2 is on the development of a Federated Learning (FL) framework, which is based on the availability of data distributed in different locations; T3.3 deals with the development of multimodal audio-visual AI models, and T3.4 deals with the establishment of an efficient distributed E2F2C Machine Learning (ML) model deployment. The analysis on the intersection of the data/use cases/models in this document guides these tasks in what is required for the implementation of the pilots.

The objective in WP5 is to ensure successful delivery of the E2F2C framework that allows for scalable and real-time processing of extreme-scale multimodal data on top of the distributed deployment of the ML models.

D8.2 is a document that specifies the MARVEL Data Management Plan and provides details on how the data will become FAIR, the data value chain, ownership, access control rights, accessibility, technical, organisational, legal and commercial aspects of data sharing to the consortium. Some of these aspects are considered in this document.

2 Use Cases, AI Tasks, and AI-Based Components

The type of data to be captured and annotated is motivated by the use cases showcased in the pilots. This section, therefore, starts with a description of the use cases and ends with a high-level description of AI tasks and AI-based components that are potentially useful and necessary to implement the use cases. The AI tasks and components will be mostly trained and tested, first on publicly available datasets and then on the MARVEL Training Datasets.

2.1 Use Cases

This section describes ten use cases (Table 1) that will be implemented primarily in the Municipalities of Trento and Malta, (MT and GRN respectively) while two of them will be developed and tested in a simulated urban environment in Novi Sad (UNS). The descriptions include pointers to the AI tasks that are likely required to implement the use cases. In addition, some use cases across municipalities are similar in nature thus providing space for the testing of federated learning.

Table 1. Use cases planned for implementation

Section	Use case	Location	Responsible Partner
2.1.1	Safer Roads	Malta	GRN
2.1.2	Road User Behaviour	Malta	GRN
2.1.3	Traffic Conditions and Anomalous Events	Malta	GRN
2.1.4	Junction Traffic Trajectory Collection	Malta	GRN
2.1.5	Monitoring of Crowded Areas	Trento	MT
2.1.6	Detecting Criminal and Anti-Social Behaviours	Trento	MT
2.1.7	Monitoring of Parking Places	Trento	MT
2.1.8	Analysis of a Specific Area	Trento	MT
2.1.9	Drone Experiment	Novi Sad	UNS
2.1.10	Audio-Visual Emotion Recognition	Novi Sad	UNS

2.1.1 Use case I – Safer Roads

This use case addresses the need to increase safety on urban roads for vulnerable road users, with the aim of encouraging the uptake of active travel modes in Malta. More specifically this use case targets cycling. Malta has witnessed a significant effort, from both the authorities and bicycle commuting lobby, in encouraging cycling and walking, mainly through infrastructural changes. The use case takes this effort further and aims at detecting cyclists, including E-bikes and possibly other motorised micro-mobility modes exiting a junction and alert the car and motorised-vehicle drivers of their presence via variable message boards in the hope that car drivers take greater care and concentrate more in such circumstances. In addition, detecting vulnerable road users is a particularly interesting task in low-visibility because it is more dangerous for these entities and it is more challenging from a technology point of view.

From an AI task point of view, this use case requires traffic entity (cycles, pedestrians, cars, etc) detectors which are typically installed at a junction. It is also necessary to resolve the exit carriageway taken by the vulnerable road users such that the respective message boards on that

carriageway are triggered, avoiding false positives, the occurrence of which can reduce the system's impact in the long term.

In addition, it would be interesting to study whether driver behaviour improves when such a system is implemented. To do so it is necessary to monitor a set of motorised vehicle variables, such as speed along the stretch of road that is used by the detected vulnerable road users. Invariably, this necessitates the temporary installation of an additional data collection system that records variables such as speed and vehicle trajectory along the road.

The detection and classification of entities is typically implemented using computer vision techniques. Detecting the cyclist is a known hard problem and at face value the addition of the audio signal would not help. However, sound signals may potentially disambiguate a bicycle from a motor cycle or moped. In addition, audio-visual models may differentiate between bicycles and motorised bicycles, which is a desired function in use case-IV.

2.1.2 Use case II – Road User Behaviour

This use case addresses the need to monitor the behaviour of road users at a junction. A potential useful application of this use case is in education campaigns targeting responsible driving, cycling, and other actions on the road. Malta has experienced fast changes in the transport landscape to which human response often lags technical progress. Educational campaigns are one way for closing the gap. The use case involves the classification of actions into a spectrum of examples demonstrating good to bad behaviour. The output can be used in the evaluation of education campaigns, which can take many forms, such as paid adverts on various media, including social media, but also the installation of information boards at targeted zones. Examples of actions at junctions include, the way pedestrians cross over the intended crossings, whether cyclists dismount at pedestrian crossings and whether car drivers stop in the delineated zone at junctions.

At a high-level, the system requires models that take as input multi-modal data (audio-visual streams) and output a summary of interesting AV segments. At a lower level, desired models need to compute with approximate speed, detect anomalous sounds (like vehicle horns, bicycle bells, excessive speed, etc), work out the instantaneous location of an entity, and track the trajectory of the entities across the junction.

2.1.3 Use case III - Traffic Conditions and Anomalous Events

This use case is on monitoring traffic conditions and detecting anomalous events, for example, traffic jam, accidents, car stuck in the middle of the junction, very slow vehicle and service vehicles parked on the side or obstructing the carriage-way. For example, the latter event is a frequent occurrence in Malta's narrow one-way urban streets, often causing ripple effects in the immediate area. In general, the output finds application in systems that are intended to inform the drivers downstream of the detected anomaly or to infer possible upstream issues in non-monitored areas, thus informing drivers of obstacles that lie ahead. In addition, the detection of anomalous events alerts personnel stationed at the traffic management control room, who can then interpret the data and take the necessary action.

From a model point of view, this use case requires the detection of anomalous data and how data deviates from the normal. Without loosing generality, anomalous events can be detected (a) in space: a number of standstill cars due to either an accident or engine breakdown or stationary service vehicle, whilst the other vehicles slow down to go around the obstacle), and (b) in time: anomalous low traffic speed or queue lengths.

One way of approaching this task is to track the vehicles over the junction/road segment and obtain an estimate of speed over time, which time-domain numerical data are then used in an anomaly detection setup. An alternative way is to build models that output discrete labels of traffic conditions, taking as input a short audio-video stream of constant time duration. The former system requires the entity detection and tracking dataset common to other use cases, whilst the latter requires a different dataset. The attractiveness of the latter lies in the fact that it may be possible to perform the processing solely at the edge.

2.1.4 Use case IV – Junction Traffic Trajectory Collection

This use case is focused on the requirement of long-term data analytics that shed light on both the behaviour of road users (e.g., car drivers, motorcyclists, cyclists, pedestrians, etc.) and on gathering traffic statistics at road network junctions. This use case is of interest for long-term transport planning and evaluation. In particular, there is currently significant interest in studying active travel modes, such as cycling and walking and in general micro-mobility. Authorities in Malta are interested in, for example, finding the optimal position of a pedestrian crossing and whether provisions for cyclists at a round-about junction are adequate and whether the installed provisions are being used as it was intended.

This use case requires entity detection and its trajectory across a junction or road segment and descriptive statistics of network junction traffic. It, therefore, follows that entity detection and tracking models are potentially used as a first processing stage followed by further processing to generate descriptive statistics.

2.1.5 Use case V - Monitoring of crowded areas

The goal is to select views of relevant areas for reasons such as exceptional crowd, suspect or unusual crowd movements, etc. The selected view could be possibly augmented with other information. The crowd-counting AI models can be a good fit in this scenario.

One of the possible areas for this scenario can be a square hosting the “Christmas Markets” (located in Piazza Fiera, Trento). Every year, from November till the first day of January, in Trento, some of the main squares of the city host the latter activities which are visited by thousands of people during the opening period. Particularly during the weekend and holidays, these areas are highly crowded. Due to these circumstances, the number of robberies and aggressions may increase. In addition, first aid may be needed for people who are unwell or faint. In order to prevent these actions, thanks to the cameras already installed and the microphones that will be installed in the square, the MARVEL framework will be adopted to potentially prevent these actions. Upon detection of such event, an alert is sent to the local police operational centre or in a control room managed by the local police.

Another potentially good candidate is the weekly market located in the city centre (located in and near Piazza Duomo). This is also a scenario where crowding can occur and where close monitoring is therefore necessary. This situation can be more challenging due to the presence of the market operators' awnings, which block most of the aerial views. The MARVEL framework will be deployed to prevent these situations by alerting the local police operational centre and consequently the policeman/woman on-site or near the market.

2.1.6 Use case VI - Detecting criminal/anti-social behaviours

The goal in this use case is to monitor areas to detect criminal or anti-social behaviours. The system triggers an alarm while delivering a custom view on the monitors in the control room. This use case will make use of anomaly detection models.

The MARVEL framework will be deployed to detect possible dangerous situations, such as gatherings, robberies, aggressions, and illicit drug trafficking, especially during the night-time. The system has to analyse the audio and visual data streams of the cameras already installed in the selected location and send an alert to the local police operational centre, who in turn can decide to send a law-enforcement squad to the location. Moreover, the audio-video stream is saved into the local server of the local police for any further investigation.

One potential location of interest is Santa Maria Maggiore Square monitored during night-time, for the detection of the presence of bothersome gangs (to detect group, noises, actions), aggressions or robberies, and gang fights.

2.1.7 Use case VII - Monitoring of parking places

This use case monitors car-parking lots. One potentially good location is the "Ex Zuffo" Parking Area which is one of the largest parking areas in Trento (around 1000 spaces) and is used by the citizens. This parking area is also effective as an interchange car park, e.g., to leave the car and reach the city centre by public transportation, rentable bikes, e-scooters.

To prevent robberies or damages to the cars parked, MARVEL framework will support prevention activities with the audio-video analysis of the existing cameras and the microphones that can be installed under the scope of the MARVEL project. Moreover, the count of cars in and out from the parking area can be integrated into the pilot, thanks to the electromagnetic traffic sensors installed under the road.

The AI tasks required in this use case include; car anomaly detection; check if only taxis park in the taxi rank (requires the discrimination of taxis from other vehicles); check the disabled parking spaces; check number of campers (requires the inclusion of new types of vehicles); average length of stay (requires post-processing of AI models).

2.1.8 Use case VIII - Analysis of a specific area.

This use case entails the collection of data (number of persons, cars, trajectories, events) in a specific area of the town (train station, schools) to support long-term decision-making by public authorities.

The Municipality of Trento wants to monitor the city's main places to support the Administration's decision-making. In order to do this, MARVEL framework can help with the counting of persons, cars, buses, taxis, bike, calculate their trajectories and calculate any notable event during a specific timeframe (for example, from 7.30 to 8.30 for school or 17.00 to 19.00 for the train station) or the entire day.

Moreover, the count of cars in and out from the parking area can be integrated into the pilot, thanks to the electromagnetic traffic sensors installed under the road, the usage of e-scooters, and e-bikes rental services. The target area for this use case is the Trento train station on the side of Piazza Dante (from Via (street) Dogana traffic lights to Via (street) Pozzo traffic lights).

This use case will mostly make use of object detection and tracking models, but also anomaly detection and crowd counting models.

2.1.9 Use case IX – Drone Experiment

Within drone experiment, we want to evaluate the potential of drones in the monitoring of large public events. To perform surveillance and monitoring of such events where either the problem is the lack of infrastructure or crowded frontal views, we can utilise a drone equipped with a camera, additional microphones, and computational resources. The drone needs to fly over the main event points to check if some problem has occurred. If the camera on the drone spots a

problem, the drone can move closer to the identified location or inform the event organisers about the problem occurrence. The drone will be utilised for aerial data capturing, but at the same time, we will explore the possibility to capture additional data from the ground using devices such as smartphones through AUD's sensMiner app (Section 3.1.1). In the experiments with staged recordings, we will also use additional microphones on the ground.

The focus of the use case is also on federated learning and distributed processing onboard. The idea of the pilot is to perform crowd classification; a working example can be to discern among the three classes of crowd behaviour, Neutral, Party, and Anomalous/dangerous behaviour. Later, classes can be further refined/multiplied as needed.

In terms of data capturing devices, the experiment will involve: several microphones on the ground, one or two microphones on board the drone, and a drone-mounted camera; we will also consider installing additional camera/s on the ground.

The use case is off-line (i.e., data stored and processing off-line), with possible real-time proof-of-concept for distributed onboard processing and autonomous decision.

2.1.10 Use case X – Audio-visual emotion recognition

The idea is to perform audio-visual emotion detection from a close-up camera and a microphone. The task will be to distinguish between several basic emotional states: neutral, happy, angry, sad, and scared, that are recognised in the psychological sciences as universally experienced in all human cultures (Jack et al, 2014), and as the use case evolves the list of emotions could potentially be increased to account for additional emotions of particular interest to the project. Detection will be based on acoustic features extracted from the audio signal (e.g., fundamental frequency, MFCCs, energy, etc.) and positions of specific points of human face (e.g., mouth corners, eyebrows, etc.) automatically detected on the video.

Audio and video are complementary to each other; sometimes both of the modalities are available for processing, but not always. Facial expressions usually clearly express emotions – like smile for happiness and frowning for anger. However, based on the angle of recording or even just the level of expressivity, facial expressions do not have to be clearly visible. Audio signal can help since it is proved that voice characteristics greatly depend on the underlying emotional state. For example, happiness will be expressed by higher pitch and usually faster speech, while sadness is typically expressed by slower speech. Some accompanying expressions like laughter, cry, inhales or similar, can also help in distinguishing emotions.

Automatic detection of emotions can prevent unexpected situations. For example, detecting fear on someone's face in the crowd (in subway, market or shopping mall) can indicate that something is wrong – the person can be lost, or even kidnapped. Anger on someone's face can indicate that the person can be the cause of some fight or similar action. Also, such scenario can for example prevent a bank to let the burglar in.

2.2 AI Tasks

The AI tasks that makeup the MARVEL audio, visual and multimodal AI subsystem are described in this section. These tasks are necessary in the implementation of the use cases and are selected from domains in which partners have expertise. The descriptions of the AI tasks are therefore organised per partner (AU, TAU, AUD). Table 2 summarises the AI tasks, which are later tabulated again in the analysis section (section 5.1) and matched with the use cases and datasets.

Table 2. List of AI tasks considered for implementation

Section	AI Task Description	Contributing Partner
2.2.1.1	Visual Anomaly Detection	AU
2.2.1.2	Audiovisual Anomaly Detection	AU
2.2.1.3	Visual Crowd Counting	AU
2.2.1.4	Audiovisual Crowd Counting	AU
2.2.2.1	Acoustic Scene Classification	TAU
2.2.2.2	Sound Event Detection and Sound Event Localisation and Detection.	TAU
2.2.2.3	Automated Audio captioning	TAU
2.2.3	openSMILE (devAIce)	AUD

2.2.1 AI tasks from AU

AI tasks that AU will consider are focused on visual and multimodal (audio-visual) scene analysis and will be examined and developed during WP3, WP4, WP5 and WP6. These tasks employ Deep Learning (DL) methods and visual or audio-visual datasets. The DL methods analysing visual information take as input an image or video and output the visual analysis result, which can be a number, an output image (density map), or a bounding box. DL methods analysing audio-visual information take as input an image or video and audio signal and output similar results for the different types of inputs. The tasks are two and they are approached by considering either only visual or audio-visual information as input. Specifically, the tasks are visual anomaly detection, audio-visual anomaly detection, visual crowd counting, and audio-visual crowd counting. Each of these tasks is described in the following sections.

2.2.1.1 Visual anomaly detection

The goal of visual anomaly detection is to establish a representation of "normal situation" in a scene based on the available training data depicting normal situations in that scene over time, and detect whenever an event occurs that sufficiently deviates from such normal situations. For instance, for a camera overlooking a pedestrian walkway, any non-pedestrian objects would be considered anomalous, and for a camera overlooking a street, any illegal and/or non-frequent activity such as jaywalking is anomalous. It is important to note that the anomalies flagged in the dataset are excluded during the training process and only used to evaluate the performance of trained models. Two typical datasets for this task are the UCSD Pedestrian dataset (Mahadevan et. al, 2010) and the Street Scene dataset (Ramachandra & Jones, 2020). This task can be employed in all use cases that will be considered in MARVEL.

2.2.1.2 Audio-visual anomaly detection

Similar to video anomaly detection, the aim of audio-visual anomaly detection is to establish a representation of "normal situation" based on both audio and visual information available in the training data, and detect whenever an event occurs that sufficiently deviates from normal situations. For instance, in a scene recorded in a train station, an anomalous video clip could depict people running away from something and the corresponding audio signal could contain gunshot sounds. To the best of our knowledge, there are no publicly available datasets for audio-visual anomaly detection in crowds, however, one approach used in the literature (Rehman et al, 2021) is to combine an audio anomaly dataset with a video one to manufacture an audio-

visual dataset. This task can be employed in all use cases that will be considered in MARVEL, where audio and visual information will be collected simultaneously.

2.2.1.3 Visual crowd counting

Crowd counting is the task of counting the total number of people present in an image. The images could contain very few people, hundreds or thousands of people (for instance, in stadiums) or even no people at all (background-only images) and may be taken from various perspectives and at different times of the day or at night. In most methods available in the literature, the output of the model is not just a single number representing the total count, but a density map that shows the number of people at different locations in the image, where the total count would be the sum of all locations in the density map. Two typical datasets for this task are the Shanghai Tech dataset[1] (Zhang et al, 2016) and the World Expo '10 dataset[2] (Zhang et al 2015). This task can be employed in the use cases dealing with crowd counting, i.e., those in the municipality of Trento and the city of Novi Sad.

2.2.1.4 Audio-visual crowd counting

Similar to visual crowd counting, the goal of audio-visual crowd counting is to count the total number of people present in an image. However, in this case, each image is accompanied by an audio clip corresponding to the ambient noise in the scene where the image was taken. The ambient audio has been shown to improve the accuracy of crowd counting in situations where the quality of the image is low, for instance, low resolution, low illumination, severe occlusion or presence of equipment noise. Similarly, to the visual crowd counting task, the output of audio-visual crowd counting models is a density map representing the number of people at different locations of the input image. The only publicly available dataset for this task is the DISCO dataset[3] (Hu et al 2020). This task can be employed in the use cases dealing with crowd counting and audio and visual information will be collected simultaneously, for example, those in the municipality of Trento and the city of Novi Sad.

2.2.2 AI tasks from TAU

AI tasks that TAU will consider are focused on situational awareness using audio signals and will be examined during WP3, WP4, WP5, and WP6. These tasks employ DL methods and audio-based datasets. The DL methods take as an input an audio signal and output different information regarding the contents of the input audio signal. The tasks are three and they exhibit some overlap regarding the exploited information in the audio signal. Specifically, the tasks are acoustic scene classification (ASC), sound event detection/sound event localisation and detection (SED/SELD), and automated audio captioning (AAC). Each of these tasks is described in the following sections.

2.2.2.1 Acoustic scene classification

ASC is the task of recognising the acoustic scene of an audio signal, that is where the recording of the acoustic signal was performed (e.g., office, street, bus). In a typical scenario, the training of an ASC method requires a dataset consisting of audio signals where each signal is annotated with the acoustic scene of its recording (e.g., DCASE ASC datasets[4]).

[1] https://svip-lab.github.io/datasets.html

[2] http://www.ee.cuhk.edu.hk/~xgwang/expo.html

[3] https://dtaoo.github.io/dataset.html

[4] http://dcase.community/challenge2018/task-acoustic-scene-classification

The task of ASC can be employed in all use cases that will be considered in MARVEL. Though, the information gained from ASC can be considered redundant in the use cases characterised by the static placement of microphones and an acoustic scene that does not change over time. However, in use cases where the microphones are moving (e.g., because the microphones are mounted on a drone or mobile phone) or the acoustic scene changes (e.g., from "heavy traffic" to "light traffic"), then ASC can be used to indicate the acoustic scene.

2.2.2.2 Sound event detection/sound event localisation and detection

SED/SELD is the task of detecting sound events in an audio signal (SED) and localise them (SELD) with point of reference the location of the microphone that did the recording of the audio signal. In a typical scenario of SED/SELD, a method takes as an input an audio signal and outputs which of some predefined sound events are active, in a pre-determined unit of time. If localisation is also performed, then the relative (to the point of recording the audio signal) location is also indicated (e.g., with azimuth and elevation). Thus, the dataset for SED/SELD needs to have the predefined sound events annotated, for the predefined unit of time. If SELD is performed, then the relative location (with respect to the point of recording of the audio signal, i.e., the microphone) needs to be annotated for each of the active sound events.

Again, the task of SED/SELD is relevant to all use cases, as it will provide indication for events happening in the monitored areas.

2.2.2.3 Automated audio captioning

AAC is the task of automatically creating natural language descriptions for the contents of an audio signal. It is not speech-to-text, as AAC does not transcribe speech ("Two people talking", AAC, vs "Hi, how are you? Good, and you?", speech-to-text). In a typical scenario for AAC, a method takes as an input an audio signal and generates a textual description (i.e., a caption) for the contents of the audio signal. The generated captions can contain information about the sound events happening in the input audio signal, the acoustic scene of the audio signal, spatiotemporal information regarding the relevant position of the sound events (e.g., far or near, foreground or background) and their interactions (e.g., before, after, or while), and information about high-level knowledge (e.g., counting, description of surroundings, description of textures). For example, a caption could be "A vehicle glass repeatedly hit a metal rod, breaking glass with multiple strikes, then the rod drops to the ground"[5].

The task of AAC could be employed in the use cases dealing with crowd monitoring, for example in the municipality of Trento or in the city of Novi Sad. The output of the AAC can provide information about the activities taking place in the monitored areas, allowing for further processing of the information, discovery of interactions between the events in the monitored area, and mining of high-level knowledge.

2.2.3 The openSMILE (devAIce) Tool from AUD

AUD's devAIce framework, of which the openSMILE toolkit is but one internal component, can be used for data processing on several computational nodes. Both devAIce and openSMILE are written in C++. On edge devices with limited computational capabilities, a trimmed-down version containing just the openSMILE toolkit can be used instead. Otherwise, if the computational node supports it, the devAIce framework will be used as a high-level wrapper. In what follows, we will refer to openSMILE as the data collection tool, assuming it will be used inside devAIce whenever necessary.

[5] Actual caption from the Clotho dataset.

The openSMILE[6] toolkit is a state-of-the-art feature extraction tool for audio streams. It supports several standard feature extraction algorithms (e.g., spectrograms, MFCCs, etc.). The specific features and hyperparameters (e.g., window size for spectrogram) thereof are configured by the user in a configuration file. openSMILE accepts a raw audio stream as input (uncompressed PCM in either 32-bit float or 16-bit signed integer format) and returns a stream of features conforming to the configuration specifications.

The owner of the data collected is the data provider using openSMILE on its premises. openSMILE works locally and does not stream any data or metadata to AUD's infrastructure. The data provider is responsible for collecting and storing the output feature stream(s), including the propagation of those streams outside the computation node running openSMILE. For example, if openSMILE is executed on a microphone device, and the features need to be transmitted to an upstream server for storage or further processing, the data provider (or, otherwise, the owner of the computation node) needs to transmit those features by themselves.

Several feature extraction algorithms supported by openSMILE are **not** privacy-preserving. For example, spectrograms contain information both about speaker identity and speech content, thus infringing upon the privacy of individuals. Privacy considerations need to be handled outside of openSMILE, e.g., either by anonymising the audio streams beforehand, or anonymising the audio features afterwards. Section 6.2 contains more information on data anonymisation.

2.3 AI-based components

The AI-based components are additional modules that are necessary in the implementation of the MARVEL framework and are based on AI algorithms. The components are selected from domains in which partners have expertise and the following sections are organised as per partner (FBK, UNS). Table 3 summarises the AI-based components, which are later tabulated again in the analysis section (section 5.1) and matched with the use cases.

Table 3. List of AI-based components

Section	AI-Based Component Description	Contributing Partner
2.3.1	Federated Learning Framework	UNS
2.3.2.1	Sound Event Detection at the Edge	FBK
2.3.2.2	Video Anonymisation	FBK
2.3.2.3	Audio Anonymisation	FBK

2.3.1 AI-based component from UNS – Federated Learning

UNS will work on several AI tasks, including audio and video event classification as a part of multimodal classification. However, the main AI task of UNS will be the development and realisation of a Federated Learning framework (FedL component) according to WP3, T3.2, in GA.

This will be achieved by distributing the training tasks to all entities of the framework, starting from the cloud all the way to edge devices. This AI task is potentially applicable to all use cases within the project as this is a general framework for learning enhancement. Typically, datasets over which FedL will run will be naturally distributed among several partners. However, in

[6] Complete documentation can be found in https://audeering.github.io/opensmile/

early stages of the project, UNS will consider performing a staged process too, where, for methodology development only, the same dataset is artificially divided into smaller pieces over which FL clients will run, mimicking the real-life setup.

2.3.2 AI-based components from FBK

There are three main AI tasks that FBK will contribute to, which are: sound event detection on the edge, video anonymisation, and audio anonymisation. These components are described in the following sections.

2.3.2.1 Sound Event Detection at the Edge

Sound event detection (SED) is related to WP3, in general, and T4.4 in WP4. FBK will provide and develop an edge SED kit that consists of a PCB board with embedded MEMs microphones acquiring samples at 16kHz. The processing unit is a low-power microcontroller (MCU), in which a neural network based multiclass classification algorithm runs in real-time. The current version of the kit is able to detect classes ranging from dog barking to car passing for a total of ten urban environment related classes. The model is trained with public dataset UrbanSound8k[7]. Nevertheless, the classifier can be adapted to the type of classes and acoustic environments relevant to the application domains, with the data and classes provided in MARVEL use cases where microphones can be deployed and edge processing is needed. The task fits to the following use cases of GRN: Road User Behaviour (use case II), Traffic Conditions and Anomalous Events (use case III); all MT use cases (use cases V-VIII) and the UNS Drone Experiment (use case IX).

2.3.2.2 Video Anonymisation

Video anonymisation is mainly concerning T3.1 in WP3. FBK will develop video anonymisation algorithms to remove the information that reveals the identity of a person, where the most concerning content is the faces for persons and car plates for vehicles, from the input video. The task first performs object detection, i.e., to localise all the concerning objects on the image frame, and then perform obfuscation, such as blurring the detected visual content. Regarding the obfuscation techniques, a basic blurring technique will be implemented. However, simply blurring the visual content has the drawback in deteriorating the performance of further video analytics. FBK therefore plans to develop a more advanced GAN-based face conversion technique under MARVEL to remove sensitive identity information while maintaining most contextual information untouched. Such algorithm development can be initiated with public datasets e.g., MOT[8], CelebA (Liu et al., 2015), but will require fine-tuning on the data from the use cases to perform satisfactorily. The GAN-based anonymisation technique is still an ongoing research at its early stage, thus it cannot be regarded as a developed technique, while standard techniques, e.g., blurring, are much more mature and can be deployed in a short time. It should be noted that video analytics, object detection, and GAN-based face redaction can be computational demanding to meet real-time performance. Appropriate edge devices should be chosen to host the algorithm. In general, the task can be potentially applied to use cases involving cameras. In some use cases, the task may not be necessary, where faces are small, i.e., the number of pixels is less than the minimum number, 17 by 17 reported in (Cai, Y., 2003), for humans to identify a person.

[7] https://urbansounddataset.weebly.com/urbansound8k.html

[8] https://motchallenge.net/

2.3.2.3 Audio Anonymisation

The audio anonymisation task is mainly related to T3.1 and T4.2. FBK will develop algorithms to remove information about the speaker identity from an audio stream. The tool could possibly replace other more invasive privacy-preserving strategies (i.e., removing segments with speech content) or approaches that rely on extracting and anonymising features. This goal can be achieved by adopting state-of-the-art approaches based on McAdams coefficients or using articulated pipelines including ASR, xvector modification, and TTS, as described in the Voice Privacy Challenge[9]. The underlying idea is to adopt GAN or VAE-based voice-conversion techniques to make the voice features in the sound unrelated to those of the actual speaker. FBK will exploit public benchmarks, e.g., Librispeech[10], VoxCeleb[11], for algorithm development and testing. It is to be noted that training and testing data is needed to check the performance in the use cases, for what concerns outdoor environments and related interfering noises. A relevant aspect to investigate is how these algorithms affect the following processing stages (i.e., SED). Potentially, the audio anonymisation would fit all the use cases. However, in some cases, it may be not necessary if no speech content is audible.

[9] https://www.voiceprivacychallenge.org

[10] https://www.openslr.org/12

[11] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/

3 Data Capturing, Annotation and Experiment Design

This section describes the tools, devices, and experimental setups used to collect and annotate data mostly for the purpose of model training development, i.e., historical data that is used to collate and develop the initial MARVEL datasets. These datasets are built from both real-life recordings as well as from staged experiments, in which actors are engaged. It also describes the plans for the methods and equipment that will be used to collect and stream real-life data that will be potentially used in the augmentation of the datasets and in the execution of the pilots.

3.1 Data Collection Tools and Devices

This section describes the data collection tools and devices provided by MARVEL partners, i.e., AUD's SensMiner tool, and IFAG's MEMS technology microphones, and how these tools and devices can be used and integrated in the various pilot campaigns.

3.1.1 The SensMiner Tool

SensMiner[12] is a data collection app and only works on Android phones (min. SDK version 21). SensMiner collects raw audio data, GPS information, timestamps, and user-defined tags. Audio is recorded in 16bit stereo PCM format in 44100Hz. Audio information and metadata are stored on the device. As such, it is under the ownership of the data providers, or the owners of the specific device. The data needs to be manually exported from the smartphone for further processing (e.g., copying it via USB to an external hard drive). As such, it is recommended to use it only during the initial data recording phase for collecting training data.

The metadata are stored in JSON format. A hypothetical example looks as follows:

```
{ "id" : "2020-06-23_112318", "startTime" : 1592904197702, "endTime" : 1592904202103, "expCode" : "marvel",
"situation" : { "id" : null, "conditions" : "partying", "lastUsageTimestamp" : 0, "recordingContext" : "Crowd",
"btaddress" : "", "btname" : "", "bttype" : "", "bluetooth" : false }}
```

The id field is then used to associate metadata with the audio file.

As SensMiner is a data collection tool that collects raw audio data, it is by definition not privacy-preserving. The data needs to be properly anonymised before being shared with other partners or the outside world.

3.1.2 MEMS Microphones and Supporting Hardware

The IFAG microphone, **IM69D130**, Figure 1, is designed for applications where low self-noise (high Signal-to-Noise Ratio or SNR), wide dynamic range, low distortions, and a high acoustic overload point is required. Infineon's Dual Backplate MEMS technology is based on a miniaturised symmetrical microphone design, similarly as utilised in studio condenser microphones, and results in high linearity of the output signal within a dynamic range of 105dB.

The microphone distortion does not exceed 1% even at sound pressure levels of 128dBSPL. The flat frequency response (28Hz low-frequency roll-off) and tight manufacturing tolerance result in close phase matching of the microphones, which is important for multi-microphone (array) applications. With its low equivalent noise floor of 25dBSPL (SNR 69 decibel A-weighted or dB(A)), the microphone is no longer the limiting factor in the audio signal chain and enables higher performance of voice recognition algorithms or far field audio signal pick-up. Each IM69D130 microphone is calibrated with an advanced IFAG calibration algorithm,

[12] SensMiner is not currently available to the public, will be deployed in MARVEL (via its partner AUD).

resulting in small sensitivity tolerances (± 1dB). The phase response is tightly matched (± 2°) between microphones, in order to support beamforming applications.

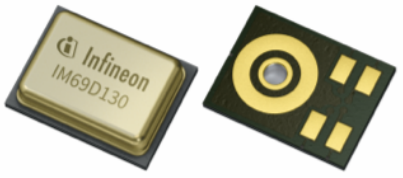

Figure 1. IM69D130

IFAG's MEMS microphones high SNR are a perfect fit with the MARVEL project goals. In free field, sound pressure halves (reduces by 6 dB) for every doubling of distance. The further the captured sound source is, the quieter the acoustic signal that reaches the microphone. As the self-noise of a microphone is practically constant, a reduction in incoming signal level causes a reduction in the SNR of the output signal of the microphone. Typically, a weak signal has to be amplified to bring it up to an appropriate level for the device signal path. Amplifying the signal also amplifies the noise present in the output. The more amplification there is, the higher the risk is that the noise will rise to a level at which it degrades the quality of the captured signal significantly. A high microphone SNR helps keep the noise floor inaudible even when the signal is amplified. The longer the capturing distance, the lower the microphone self-noise should be to avoid problems. This is especially critical when the distance is long and the sound source itself is quiet. As sound pressure attenuates by 6 dB per doubling of distance, using a microphone with a 6 dB higher SNR can enable doubling the cap. For more detailed technical information please refer to the datasheet[13].

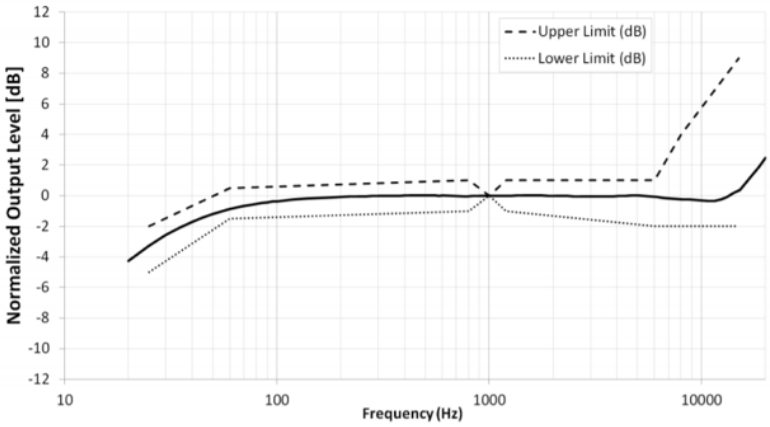

Figure 2. IM69D130 free field frequency response

IFAG has provided two boards that enable the data providers in the gathering of data, Infineon's Audiohub Nano and Infineon's Shield2Go MEMSMIC.

The IFAG **Audiohub Nano** (datasheet[14]) enables the evaluation of Infineon digital pulse density modulation (PDM) XENSIV™ MEMS microphones. Up to two IFAG digital XENSIV™ MEMS microphones can be connected to the evaluation board in mono or stereo

[13] https://www.infineon.com/cms/en/product/sensor/mems-microphones/mems-microphones-for-consumer/im69d130/

[14] https://www.infineon.com/cms/en/product/evaluation-boards/eval-ahnb-im69d130v01/

output. The evaluation board provides a USB audio interface to stream audio data from the microphone with any audio recording and editing software. Some features available in this board are:

- Audio streaming over USB interface
- 48 kHz sampling rate
- 24-bit audio data (stereo)
- Mode switch for toggling between normal mode and low power mode with 4 pre-defined gain configurations
- LEDs indication for the configured gain level in normal mode and low power mode
- Volume unit meter display with onboard LEDs
- Powered through Micro-USB

IFAG's **Shield2Go MEMSMIC** (datasheet[15]) boards offer a unique customer and evaluation experience – the boards are equipped with two High-performance digital MEMS Microphone IM69D130 and come with a ready to use Arduino library. Data providers can develop their own system solutions by combining Shield2Go boards together with external hardware solutions like Arduino and Raspberry PI, which are popular hardware platforms. All this enables the fastest evaluation and development of the recording system. This boar offers features like:

- 2x IM69D130 Digital MEMS microphone in stereo mode configuration
- PDM and I2S output configuration
- Flexibility to develop a custom application with Arduino and Raspberry PI

The data provided by the microphones/boards is raw audio data in different formats (PDM, I2C, or USB audio format). Neither the Microphone nor the boards store the audio data, it is streamed to the node in charge of the audio acquisition.

When installed on the field, sealing must work well in all use cases throughout the device lifetime. The device implementation must also provide the microphones with proper protection from environmental conditions such as dust and liquids as well as abuse such as impacts. The sound channel in a device is a key factor determining the performance of a microphone. It must be designed to minimise the effect of resonances. The acoustics and mechanics around the microphone must be stable to avoid short and long-term changes that could affect audio quality. Sturdy device mechanics and elimination of noise sources help prevent handling noises. Good and reliable sealing is a critical enabler for high audio capturing quality. Microphones must also be provided with a working environment free of mechanical stresses. Microphones must be placed correctly in a device in relation to the sound sources. The requirements set by multi-microphone algorithms must be taken into account. The microphones must also be placed away from noise sources and stress factors. The size of the implementation can be minimised by choosing the right microphone type for the device mechanics. A detailed installation guideline can be found in the application notes[16].

The owner of the data captured by the microphones is the owner of the node in charge of the audio acquisition, for example, the data provider. IFAG's microphones capture the raw audio data and send it to the node without any processing, i.e., any anonymisation algorithms should be implemented external to the IFAG's microphone systems.

[15] https://www.infineon.com/cms/en/product/evaluation-boards/s2go-memsmic-im69d/

[16] https://www.infineon.com/cms/en/product/sensor/mems-microphones/mems-microphones-for-consumer/im69d130/

3.2 Data Annotation Tools and Devices

This section describes iHEARu-PLAY, which is a data annotation tool provided by AUD and ELAN, which is an open-source tool developed within Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. These tools will be used in the annotation of training datasets. The former is useful to label whole short audio or video files, amongst other options, and the latter is suitable for the multi-label annotation of audio files. Both annotation types are useful in the preparation of datasets for the MARVEL project.

3.2.1 iHEARu-PLAY

iHEARu-PLAY is a web-based crowdsourcing data annotation platform. It runs on AUD's dedicated server and can be accessed through most browsers. Thus, the data has to be temporarily transferred to AUD so as to be annotated. Ownership of the data remains with the data providers, and AUD will delete all data afterwards. The completed annotations will be returned to the data providers to be included with their datasets before being shared with the consortium. By default, AUD is using its open-source data format to store annotations and metadata, and it is in this format[17] that annotations will be returned to the data providers.

iHEARu-PLAY can be used to annotate both audio and video data with categories, tags, and free text. Segmentation is currently supported for audio data only. All tasks support annotation by multiple annotators, and it is possible to specify a minimum and a maximum number of required annotations per data sample. Quality metrics (e.g., annotator agreements) can be computed by AUD after the process has finished. An example of several supported annotation interfaces for the task of emotion recognition is shown in Figure 3.

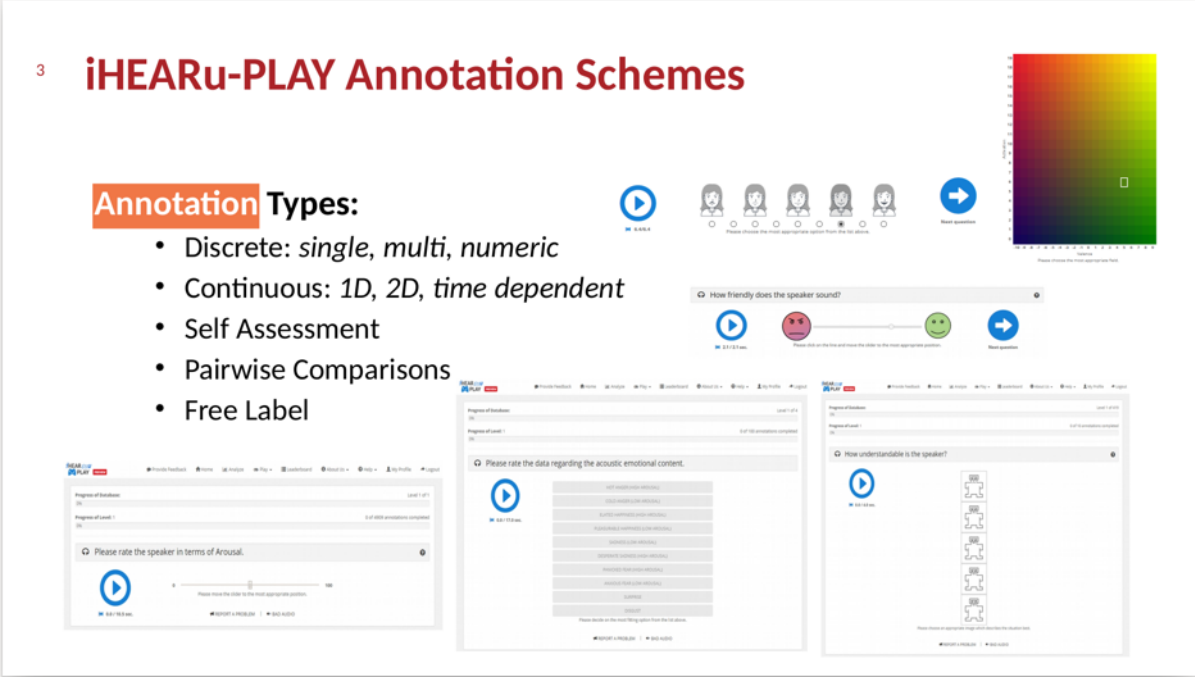

Figure 3. iHEARu-Play interface example

The platform is closed-source but can be accessed for free by authorised users through the web. User authorisation is handled by AUD and can be enabled for partners upon request. The cost of annotation corresponds to the paid work by the annotators. Annotation resources have to be

[17] https://github.com/audeering/audformat

provided by the data providers, who have to designate the persons responsible for annotation. AUD will provide authorisation to use the platform to those persons upon request, as well as access to the data to be annotated.

Naturally, annotating data potentially compromises the privacy of the recording participants. AUD is not responsible for ensuring this privacy. It is assumed that the data will be anonymised by the data providers before being transferred to AUD and that the annotators will have signed appropriate confidentiality agreements. The data will be accessible to the bare minimum of AUD personnel required to obtain the data by the data providers and make it accessible through iHEARu-PLAY. Access to all other personnel will be restricted.

3.2.2 ELAN - Data Annotation Tool

ELAN[18] is an open-source software created within Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. ELAN is an annotation tool for audio and video recordings. ELAN is useful in project MARVEL for the multi-label annotation of audio signals while viewing the time-aligned video signal in its in-built video player. For example in the annotation of traffic recordings the video feature helps the user distinguish the type of vehicle emitting a specific sound, Figure 4. During the annotation process, the annotator selects and labels a time interval. In addition, time intervals annotated with different labels can overlap each other (multi-label annotation). An example annotation file is the following:

%Label	%start_time	%end_time
Car/Brakes_Screeching	53.183	54.05
Motorcycle/Engine_Idling	86.917	100.8
Pedestrian/Voice_Chatter	90.2	111

It is also possible to segment and label the video in time, however, it is not possible to annotate the video frames with for example bounding boxes.

The ELAN tool can be downloaded[19] and used for free (requires citing in any publication on research in which ELAN has been used) and all the documentation necessary to use it is provided. The tool is customisable, making it easier to use and faster to annotate and also allows for specifying the annotation ontology. The tool executes on a local machine and does not require an internet connection. ELAN offers an option for multiple annotators by including extra tiers in the ontology. However, it may easier to collect annotations separately and compute quality metrics (inter- and intra-rater agreements) externally. Since all annotation is carried on local machines, the data providers remain the sole owners of the data and the annotations, and are solely responsible for the secure storage of the data.

[18] https://archive.mpi.nl/tla/elan

[19] https://archive.mpi.nl/tla/elan/download

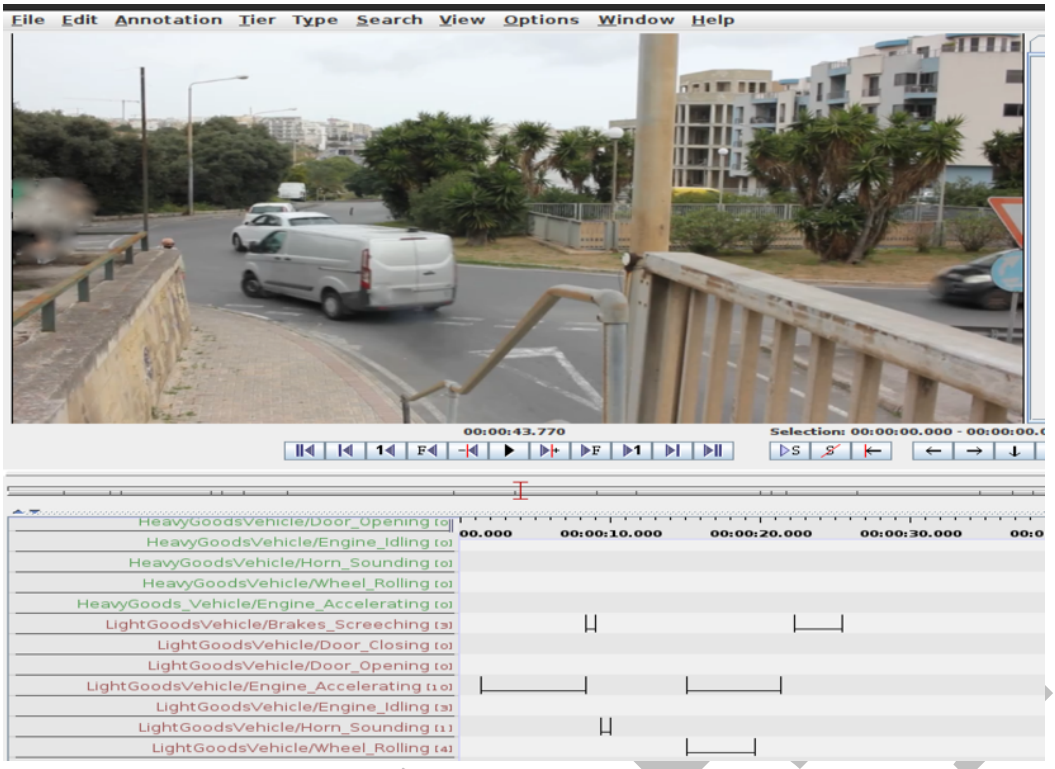

Figure 4. Example of traffic audio annotation in ELAN

3.3 Data collection for MARVEL training datasets

This section describes the equipment, methods, and experimental design used to collect data, from the municipalities of Malta and Trento (GRN and MT) and the city of Novi Sad (UNS), for the purpose of developing the initial MARVEL training datasets. Such data collection may include a mix of short real-life recordings as well as staged recordings, for which actors are recruited. The following section (3.4) describes the plans for collecting real-life recordings for MARVEL dataset augmentation and the execution of the pilots.

3.3.1 Data Collection in Malta

Two experimental setups (GRNEdge), one fully portable and the second one mobile but fixed to physical infrastructure are used to collect data for the initial training datasets. Figure 5 depicts the system data model for the collection of audio-visual data. This mobile setup is necessary to collect data of sufficient variability in terms of location characteristics, illumination, video quality, angle of view (azimuth and elevation), position of microphone sensor, and background noise that should help in the development of models that generalise better when predicting on unseen data. Data is collected at road junctions, which are defined as public spaces, and will be used to train models that detect traffic entities, such as cyclists, pedestrians, and cars, and the trajectory across the junction of the same entities. The data collected is expected to be sufficient for use cases I-IV (section 2.1). Use case I is on rendering safer roads for cyclists with the intention of encouraging active mode uptake in a municipality where very few commute on bicycles and it will probably be necessary to recruit volunteers to cycle during specific data collection periods. This necessitates obtaining consent from the volunteers (consent form will be provided in D9.1 due M12). The rest of the data collection is carried out in the wild. All PD will be removed from the audio-video data and no PD will be included in the metadata and textual annotations.

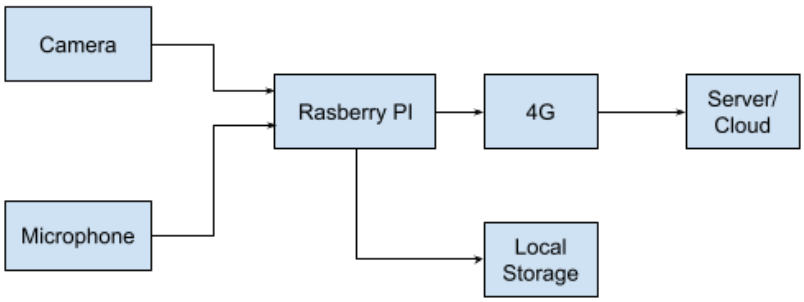

Figure 5. The experimental audio-visual data collection setup (GRNEdge)

A Raspberry Pi Camera Module 8MP v2.1 is used for capturing the video and the IFAG MEMS microphone is used to capture the audio at 44.1kHz sampling rate and 24-bits per sample. The video camera's maximum resolution is 1920x1080 at a frame rate of 30fps. Synchronisation of the two data streams and their encoding takes place on the Raspberry Pi board. The average encoded data rate is estimated at 12Mbps at full video resolution and 30fps down to 1.5Mbps when down-sampled to 640x480, @16fps. The synchronised audio-video is delivered in MKV format. The variety of data is audio and video, the velocity varies from 16fps to 30fps per camera/audio system and the volume generated varies from 0.75GB/hour to 5.5GB/hour of audio-video data. This data is either stored on local storage at the edge or transmitted on a secure channel to a local server. The schema for the metadata is as follows:

[Setup ID], [Location: GPS coordinates, street address, date, time], [Details on camera: model number, lens, focal length, height, look angle] [Microphone: model number, mono/stereo, type, direction], [Video: frame resolution, frame rate, encoding format], [Audio: sampling rate, quantisation, encoding format]

Following data collection, approximately 4 hours of audio-video data consisting of short clips of 4-5 minute duration are selected. Anonymisation of the short clips is carried by masking faces and vehicle number plates in the video stream and human speech in the audio stream, as described in section (6.1). The anonymisation stage is followed by the annotation process for which the software tool ELAN (section 3.2.2) will probably be used.

3.3.2 Data Collection in Trento

Experimental setups in Trento aim to monitor selected public spaces and inform the control room of local police if something anomalous happens: some public places of the city centre of Trento will be used to detect possible dangerous situations such as gatherings, robberies, aggressions/fights, drug dealing, car accidents, etc. Events will be notified to the central station via an alarm and by creating a custom view on a smart interface to highlight the relevant cameras (Figure 6). Events will also be saved for further analysis.

The general goals described above can be mapped into four specific use cases; monitoring of crowds, detecting criminal/anti-social behaviours, monitoring of parking places, and analysis of a specific area. Real recording includes data acquired by the surveillance cameras currently mounted in four sites. Currently, the cameras generate data at the rate of one image per second (1fps) per camera. Microphones (IFAG microphone arrays) will be mounted nearby the cameras, but speech has to be removed or made inaudible (anonymisation process). Data acquisition from the microphones and storage in the data centre are open issues yet to be solved.

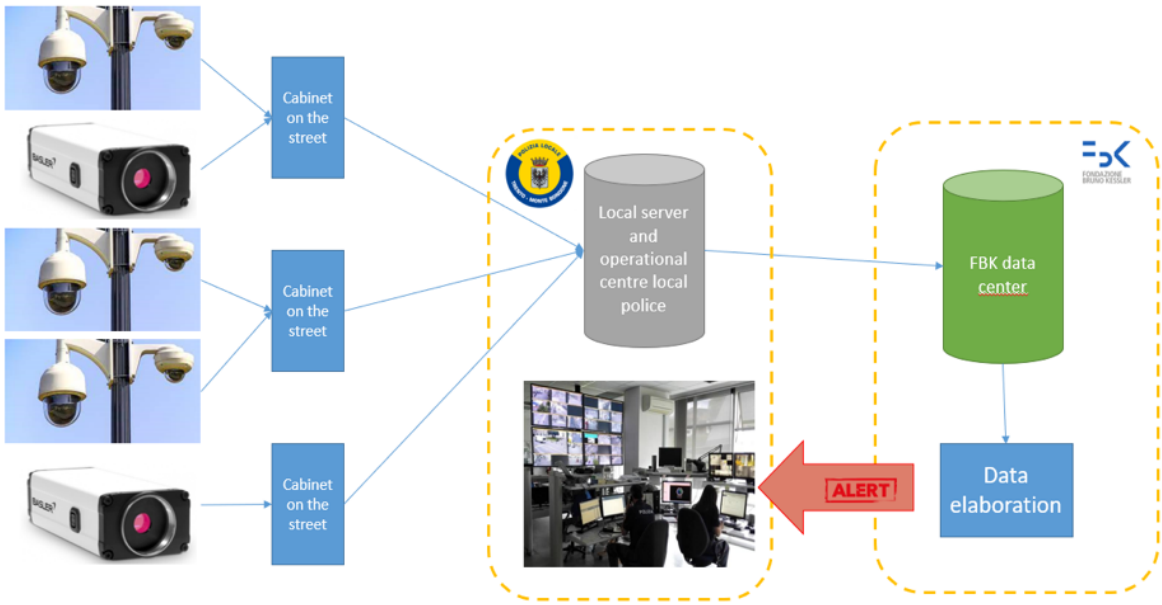

Figure 6. Infrastructural plan for the collection and processing of data in the MT pilot

In order to complement the real data, in particular for what concerns rare events and their labelling and to tackle the limitations due to privacy, MT and FBK will record target events staged by enrolling participants from the two institutions. Data will be recorded with: three cameras, three IFAG MEMS arrays, and three large aperture ST MEMS arrays. Recordings will probably take place in a different location with respect to the use case sites. Different light conditions will be considered: daylight, night with artificial lights, dusk. The dataset will simulate dangerous situations and isolated events related to the above dangerous situations. Below more details of the datasets are provided.

The TrentoOutdoor - **Real dataset** is suitable for all the trial cases planned, i.e., monitoring of crowds, detecting criminal/anti-social behaviours, monitoring of subway, monitoring of parking places, and analysis of a specific area. The data is delivered as unstructured raw video and audio data (mp4 and wav) and includes pdf documents for description and annotations that will be used to train/develop and evaluate the AI algorithms. Other public datasets, which may help improve the algorithms, will be re-used, for example existing crowd datasets (section 4.1) for pre-training models, which can then be fine-tuned with the acquired dataset under MARVEL. Security measures will be implemented to prevent unauthorised access to personal data and MT will be compliant with DPO instructions. FBK, as data processor, will implement the standard security measures as described in section 6.2.

During the first phase, all audio and video data will be stored and anonymised (by FBK). In the second phase, only audio and video relevant to the purposes of the project will be selected. The volume of the data generated will be in the range of tens of Terabytes. Annotations will be registered through metadata in the recordings (i.e., date, time, and location weather conditions, scenario, in video stream, real or staged, day/night, etc.).

The TrentoOutdoor - **Staged Recording** dataset is necessary to simulate rare events that are otherwise difficult to be collected in real-life. This dataset will address one or two use cases. The data is delivered as unstructured and raw video and audio data (mp4 and wav) and includes pdf documents for description and annotations that will be used to train/develop and evaluate the AI algorithms. Other public datasets, which may help improve the algorithms, will be re-

used, for example existing crowd datasets (section 4.1) for pre-training models, which can then be fine-tuned with the acquired dataset under MARVEL. Security measures will be implemented to prevent unauthorised access to personal data. Informed consent from all participants will be collected beforehand and open access will be considered for all the data collected. In any case, MT will be compliant with DPO instructions. During the first phase, all audio and video data will be stored and, according to the DPO decision, the opportunity to anonymise the data (by FBK) collected will be evaluated. In the second phase, only audio and video relevant to the purposes of the project will be selected. The volume of the data generated will be in the range of tens of Terabytes. Annotations will be registered through metadata in the recordings (i.e., date, time and location weather conditions, scenario, in video stream, real or staged, day/night, etc.).

3.3.3 Experimental data Collection in Novi Sad

The UNS experiment will support the other two pilots by providing a specific small-scale use case and data for further processing. The idea of the experiment is to perform crowd classification based on drone and ground recordings. Our approach is to define a few relevant classes that will be used for crowd classification and anomaly detection. For example, the first class may be a neutral class, where people are just gathered and chatting. The second class could be a party, where people are dancing, singing and laughing. And the third class could represent some problem, where people are screaming, yelling, running, or where some anomalous behaviour, in general, is happening.

Figure 7 provides the experimental setup for the drone experiment ("Edge" and "Fog" boxes) and also serves as an illustration of the infrastructure for the corresponding Proof-of-concept demonstration ("Edge", "Fog" and "Cloud" boxes).

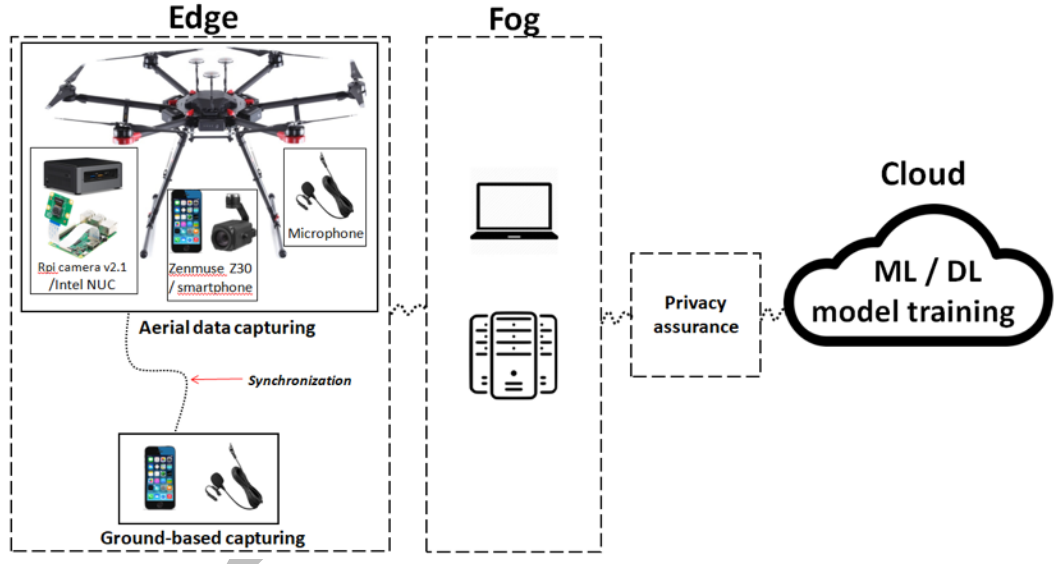

Figure 7. Experimental setup and infrastructure for the UNS Proof-of-Concept demonstration

During the experiment, the drone will mount; (a) RPi v2 Camera (possibly a DJI Zenmuse camera will be purchased), (b) MEMS microphones provided by IFAG, and (c) Processing unit(s): Raspberry PI or Intel NUC for logging and eventual edge processing. On the ground, the devices will be: (a) Camera, (b) Microphones, and (c) Mobile phone using SensMiner for labelling.

Staged recordings will be designed so as to replicate the crowd-monitoring events of interest for training. Our approach is to define a few relevant classes that will be used for crowd

classification and anomaly detection. For example, the first class may be a neutral class, where people are just gathered and chatting. The second class could be a party, where people are dancing, singing and laughing. And the third class could represent some problem, where people are screaming, yelling, running, or where some anomalous behaviour, in general, is happening. To collect examples of the identified classes, UNS will perform staged recordings with experimental participants whose written consent would be sought for the experiments' execution (will be included in D9.1).

In addition to staged recordings, real-life testing will be also performed in the form of staged pilot. However, UNS will consider other large public city events, such as festivals for possible testing if possible. If such scenario will be performed, UNS will apply all anonymisation techniques and protocols defined within the consortium.

As UNS will perform staged recordings and all participants will sign a written consent, it would not be needed to apply anonymisation techniques. However, UNS intends to contribute to the application of anonymisation techniques defined within the consortium. Anonymisation could be achieved through face anonymisation via face replacement techniques where faces of individuals recorded in camera streams will be replaced by randomly generated faces or predefined set of faces. A similar approach could be applied to audio recordings.

The UNS dataset will be created by several drone flights, each flight being approximately 15 minutes long (maximal drone hovering time with payload is around 15 minutes). As the plan is to record about 10 hours of data, at least 10 GB will be probably required for storage for medium quality. However, we believe that at least 20 GB of data or even more will be produced. The whole scenario includes drone recordings (with mounted camera), MEMS microphones, and environmental sensors to provide audio-visual footage. Images will be captured as JPEG and DNG pictures (additionally, drone can capture RAW images in Burst mode). Videos will be recorded in H.264 and H.265 codecs and saved to the Micro SD card, whereas audio data will be stored as wav files. Taking into account different scenarios, each video should be at least 5 minutes long. Audio stream will be synchronised to the video stream. To perform data annotation, the ELAN (section 3.2.2) annotations tool for audio and video recordings will likely be used.

The dataset for the audio-visual emotion recognition task will be acquired using mobile phones, through a mobile application designed for that purpose. The application will play a recorded sentence to the user with the intonation that corresponds to the target emotion. The user will have to pronounce the previously heard sentence with facial expressions that he/she feels are appropriate and will be recorded by mobile phone's camera and microphone. The heard (original) sentence (a featured example) is previously recorded by a professional actor and is part of Serbian Emotional Speech Database (Jovičić et al, 2004). The content does not necessarily correspond to the target emotion. A similar application, but recording audio content only, is already developed for the national project S-ADAPT - Speaker/Style Adaptation for Digital Voice Assistants Based on Image Processing Methods. [20]

3.4 Data augmentation through real-life streaming data

This section describes the plans that have been made for collecting real-life recordings for MARVEL dataset augmentation and the execution of the pilots. The description includes an initial indication of how the data flows through the edge/fog/cloud setup, i.e., the variety,

[20] https://www.telekom.ftn.uns.ac.rs/SADAPT.html

schema, volume, velocity, format of the data. This information will be used for the initial planning of the MARVEL data management platform, Section 5.4.

3.4.1 Real-life Data Collection in Malta

The planned setup and data model for collecting real-life data is depicted in Figure 8. This setup will be used for both the augmentation of the training dataset and for executing the pilots. The main sensors are of the type shown in Figure 8, which include cameras, microphones, and anonymisation algorithms at the edge. The anonymised AV stream (generating a volume of 5.5GB/hour/device) is live streamed to the fog layer, where the first AI models are likely to be applied. Each camera/microphone setup generates input at a constant velocity of 15-25 images or frames per second and 48k samples per second of audio signals. These inputs are encoded into an AV stream in MKV format. The fog layer converts the unstructured data into structured csv data, mainly via the entity detection and tracking AI algorithms, to generate the **GRN-TXT-traffic-data** dataset (Section 4.2.3).

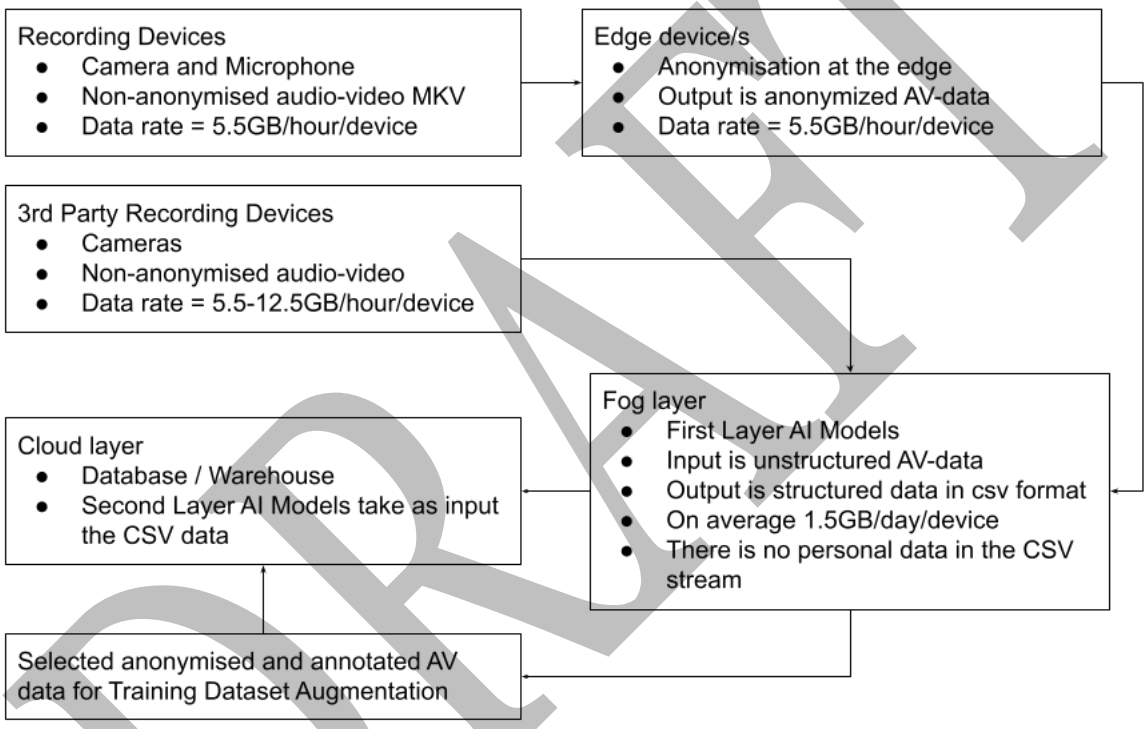

Figure 8. Proposed data model for real-life data collection in Malta

Furthermore, the fog layer extracts and stores the segments of anonymised short clips of AV unstructured data characterised by high entropy. These short AV clips are used for either augmenting the training dataset (after the appropriate annotations are added) or for further research and use in the use cases. The volume of csv data generated by the fog layer is in general stochastic in nature. Based on GRN's data collection system, it is estimated that the fog layer generates 1.5GB/day/device of structured data. The Data model in Figure 8 allows for the addition of 3rd party sensors to the system. These largely uni-modal sensors transfer data directly to a fog server, which may be, but not necessarily, co-located with other fog servers. These fog processors can generate csv output directly, by-passing anonymisation, which is only applied on the short clips extracted for training data augmentation and further research. The sources generate input at similar rates as the system in Figure 5, but potentially at lower or higher frame resolutions. The volume of AV data generated varies between 2-12.5GB/hour/device, whilst the volume of csv data is on average 1.5GB/day/device.

Both csv data and the anonymised short AV clips, which may also be annotated, are finally transferred to and stored on the cloud where the final AI algorithms are applied and where further research can be carried out.

3.4.1.1 Data Value chain and GDPR

Data is collected at the edge and transformed as it moves along the E2F2C infrastructure. The format and velocity of the data at each stage depend on the sensors as well as on the distribution of the processing algorithms over the infrastructure. Typically, the audio-visual raw data is collected and anonymised at the edge and the resulting audio-video is transferred to the fog layer where the first layer of AI models converts the binary non-structured audio-video data into structured non-binary data, which in turn is transferred to and stored in the cloud. However, it is also expected that some AI models, that take as input raw audio-video data, are executed at the edge and the resulting non-binary structured data is transferred to the fog and eventually to the cloud. In addition, part of the anonymised audio-video data stored at the fog layer is selected and manually annotated for augmenting the curated training datasets. These training datasets and the non-binary structured data are eventually added to the MARVEL Data Corpus repository.

GRN will collect audio-visual data mainly from public road network junctions. The recordings will therefore include visuals of entities that make use of roads, such as vehicles, human drivers, cyclists and pedestrians and audible sounds associated with the same entities. These audio-visual streams may therefore contain personal data. For the case of recruiting volunteers (for example cyclists) for the purpose of dataset class balancing, GRN will explain the whole process of data collection and seek the consent of the volunteers in a fully transparent manner. To adhere to GDPR and privacy regulations, the audio-visual data collected will be anonymised with a process that masks human faces, vehicle number plates, and human speech. In addition no personal data will be included in the metadata or annotations.

The owner of recorded data will be Greenroads Ltd (GRN), who will also define access control rights, legal and commercial aspects. In the eventual addition of third-party sensors the ownership, access control rights, legal and commercial aspects will be defined when the case arises.

3.4.2 Real-life Data Collection in Trento

The purpose of the data collection is to train/develop and evaluate the algorithms elaborated by the partners of the project. The generated data will be raw video and audio in mp4 and wav format respectively. In addition, some pdf documents will be added in order to have some other description and also for the annotation phase. Public datasets will be also used in order to improve even more the algorithms (e.g., existing crowd dataset for pre-training models, that can then be fine-tuned with the acquired dataset under MARVEL.

The data will be recorded from public cameras (subset of the video surveillance cameras owned by MT) and microphones (provided by IFAG). At the current stage, MT cannot calculate the size, but from a first estimation, the range would be around tens of Terabytes.

Real recording will be taken using three digital cameras already installed in each place and three IFAG mems arrays that will be installed where it is possible, not in all places of the trial cases. In order to complement the real data, in particular for what concerns rare events and their labelling and to tackle the limitations due to privacy, MT and FBK will record target events staged by enrolled participants for the two institutions. Data will be recorded with three digital cameras, three IFAG mems arrays, and possibly three large aperture ST mems arrays. Different light conditions will be considered: daylight, night with artificial lights, dusk.

At this moment all the videos are stored on a local server at the local police operational centre. In the case of the real recordings, the feeds of the cameras go straight into the data centre where images are stored and the data are then processed at FBK's servers, which are situated at the FBK premises (Figure 6). The latter act as fog layer resources, before moving the data or processed data to the cloud.

3.4.2.1 Data Value chain and GDPR

Municipality of Trento is the owner of all data collected by MT, whilst FBK has been nominated by MT as the responsible for treatment partner.

The video data is transformed to other forms of data either at the fog layer or the cloud layer, whilst for the audio data, there is the option of processing at any layer over the E2F2C infrastructure, thus including the edge.

During the first phase, all audio and video will be recorded and anonymised. In the second phase, the data will be selected and only audio and video relevant to the purposes of the project will be stored. In order to be in compliance with the GDPR regulation, Municipality of Trento will comply with DPO instructions. Municipality of Trento is a Public Administration and so it will need only privacy disclaimer from people involved.

3.4.3 Experimental data collection for dataset augmentation in Novi Sad

Experimental data collection setup for dataset augmentation and the setup for pilot execution of UNS pilot will be similar to the setup described in section 3.3.3. Dataset augmentation will be performed by recording additional data for the same classes in different conditions – daylight, evening, low light, various number of participants on the square, etc. UNS pilot is not a true pilot but the staged one. All participants will sign a written consent so that testing will be performed in a similar way. However, we will explore possibilities to perform recordings at additional locations for testing.

3.4.3.1 Data Value chain and GDPR

The UNS experiment will collect a variety of data (including audio, video, and environmental sensors) from various edge devices. In this case, there is also the added difficulty of processing data on the air-borne drone (one of the edge devices) due to payload and power limitations. In this case, it may be necessary to secure the wireless transmission of the data to a ground station, where the necessary anonymisation processing is completed before the data moves to further fog and cloud layers. As with the other two pilots, the data and associated processing tasks can be distributed according to the demands of the use case in hand and the processing and storage nodes available.

The data will be collected from experiments in a controlled environment that will involve a drone equipped with one or more cameras, microphones, and environmental sensors. Besides the data from drone-mounted sensors, the data at UNS experiment will likely include the data obtained from the multimedia feature extraction toolkit, i.e., audio recordings collected through the devAIce (AUD) technology running on the mobile phones of the human subjects participating in the experiments. For research purposes, UNS will also perform development and testing of the following MARVEL technologies: federated learning, edge ML, DL for audio-visual analytics.

In such circumstances, personal data processed will include audio recordings from microphones and mobile phones (through the devAIce app), video streams from cameras, and the age of human subjects. The research participants in the UNS experiments will likely involve pedestrians, people using mobile phones, people operating vehicles, bicycles, etc.

UNS will always obtain explicit consent from each of the human subjects participating in the respective experiment to collect and process their data. Hence, personal data that UNS will process will also include the data necessary to obtain explicit consent for experiment participation, such as names and email addresses. Transparent information about the usage of their personal data will be provided to data subjects at the time that consent is obtained and their rights with regard to their data explained, such as the right to withdraw consent. This information will be provided in an accessible form, written in clear language, and free of charge.

The owner of recorded data will be the Faculty of Technical Sciences, University of Novi Sad. This institution will define access control rights, legal and commercial aspects.

4 Datasets for Model Training

This section describes datasets that are potentially useful in the process of training the models. Section 4.1 describes open and publicly available datasets that are suited for pre-training the models, while Section 4.2 describes the MARVEL datasets that data providers will make available for the training or the fine-tuning of the models. The latter includes both labelled and unlabelled datasets. Table 4 below provides an overview of the full list of the datasets that will be used within MARVEL.

Table 4. List of datasets including both open datasets as well as datasets collected as part of the MARVEL project

Section	Name of Dataset	Type	Comments
4.1.1	MAVD traffic	Labelled Audio with Video	Public Dataset
4.1.2	UCSD Pedestrian	Labelled Video	Public Dataset
4.1.3	Street Scene	Labelled Video	Public Dataset
4.1.4	Shanghai Tech	Labelled Images	Public Dataset
4.1.5	World Expo ’10	Labelled Video	Public Dataset
4.1.6	DISCO	Labelled Audio and Video	Public Dataset
4.2.1	GRN-AV-traffic-entity	Labelled Audio with Video	MARVEL Dataset
4.2.2	GRN-AV-traffic-state	Labelled Audio and Video	MARVEL Dataset
4.2.3	GRN-TXT-traffic-data	Non-binary structured, CSV	MARVEL Dataset
4.2.4	TrentoOutdoor-Real	Labelled Video and Audio	MARVEL Dataset
4.2.5	TrentoOutdoor-Stage	Labelled Audio-Video	MARVEL Dataset
4.2.6	UNS Drone	Labelled Audio-Video	MARVEL Dataset
4.2.7	UNS Audio-Video Emotion	Labelled Audio-Video	MARVEL Dataset

4.1 Open Datasets

This section describes datasets that are publicly available and are potentially useful to train or pre-train in a first step the AI models that will be used in the implementation of MARVEL use cases.

4.1.1 MAVD[21]

The MAVD-traffic (Montevideo Audio and Video Dataset), (Zinemanas et al 2019), is a public collection of annotated road traffic sounds in the urban environment intended for the development of sound event detection models. In addition to the audio tracks, the dataset includes synchronised video recordings, which are not anonymised and therefore released in low resolution. The annotated audio-video recordings (50 in total) are approximately three to

[21] https://zenodo.org/record/3338727#.YNAPti2B1Bx

six minutes long, and add up to approximately four hours. The data is recorded at four different locations during both day and night time.

Furthermore, to facilitate annotation and allow for various level of detail, an ontology of traffic sounds is developed. The ontology is made up of a set of two taxonomies (vehicle types, e.g., "bus", and components, e.g., engine) and a set of actions (e.g., "accelerating"). The annotations, which take the form of labelled temporal-audio-segmentations are carried out manually on ELAN[22], which is an annotation tool and is described in section 3.2.2.

Finally, two base-line models (audio only) are developed for the traffic sounds, a Random Forest (RF), and a Scraper Convolutional Neural Network (S-CNN), (Salamon et. Al. 2017). The input features for the RF are Mel-Frequency Cepstral Coefficients (MFCC) computed over 40ms, and the first and second-order derivatives. For the S-CNN the input is a 1sec in length mel-spectrogram and the network is first trained on the URBAN-SED dataset and then fine-tuned on the MAVD-traffic dataset. The baseline models suffer from the unbalanced nature of the dataset and primarily do well in the detection of the majority class (car wheel rolling).

This dataset is directly related to some of the MARVEL use cases (I, II, III, IV, VII, VIII) and can be useful in (a) the training or pre-training of models, (b) influence the label annotation ontology chosen for the MARVEL traffic datasets (see section 4.2.1) to fit case studies, and (c) the available open-source code provides a baseline for the SED models in MARVEL.

4.1.2 UCSD Pedestrian dataset[23]

The UCSD Pedestrian dataset (Mahadevan et. al, 2010) for video anomaly detection consists of 98 videos at a low resolution of 158×238 or 240×360 pixels, with 120 to 200 frames per video. These videos are taken from a pedestrian walkway using a static camera from two different perspectives, and all bikers, skaters, carts, wheelchairs, and other non-pedestrian objects are considered anomalous. In the dataset, frames containing anomalies are flagged, along with the location of the anomaly in the frame. This dataset is the most widely used dataset for video anomaly detection and will be used to train our models which detect anomalies in videos and test their accuracy. This dataset is directly related to some of the MARVEL use cases (section 2.1, V, VI, and IX) and can be useful in the training of models.

4.1.3 Street Scene dataset[24]

The Street Scene dataset (Ramachandra & Jones, 2020) for video anomaly detection consists of 81 videos of high resolution of 1280×720 pixels (HD) with a total of 203,257 frames. The videos are taken from a camera overlooking a two-lane street from a single perspective and during daytime. The dataset contains a variety of naturally occurring anomalies such as jaywalking, illegal U-turns, and an officer issuing a parking ticket. In the dataset, frames containing anomalies are flagged, along with the location of the anomaly in the frame. This dataset is a recent addition to the publicly available datasets for video anomaly detection, and its advantage is the diversity of the activities within the videos as well as the high resolution and quality compared to older datasets. This dataset will be used to train and evaluate our models which detect anomalies in videos alongside the aforementioned UCSD Pedestrian dataset to ensure they are suitable in a variety of situations. This dataset is directly related to

[22] Max Planck Institute for Psycholinguistics, "ELAN -the language archive." [Online]. Available: tla.mpi.nl/tools/tla-tools/elan/

[23] http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

[24] https://paperswithcode.com/dataset/street-scene

some of the MARVEL use cases (Section 2.1, V, VI, and IX) and can be useful in the training of models.

4.1.4 Shanghai Tech dataset[25]

The Shanghai Tech dataset (Zhang et al, 2016) for crowd counting is split into two parts. Part A contains 482 images of crowds taken from the web with an average resolution of 589×868 containing a total of 241,677 people, and Part B consists of 716 images taken from a busy street in Shanghai with a resolution of 768×1024 containing a total of 88,488 people. The images in both parts are taken from various perspectives and have a variety of population densities. The location of the head of each person present in each image is annotated. This dataset is one of the largest publicly available datasets for crowd counting and is widely used within the literature. We will use this dataset to train and evaluate our models which count the total number of people present in an image. This dataset is directly related to some of the MARVEL use cases (Section 2.1, V, VI, VIII and IX) and can be useful in the training of models.

4.1.5 World Expo '10 dataset[26]

The World Expo '10 dataset (Zhang et al 2015) includes 1,132 video sequences captured by 108 surveillance cameras and thus has a wide variety in terms of perspective and population density. It contains a total of 3,920 frames with a resolution of 576×720 and 199,923 people. The location of the head of each person present in each image is annotated. This dataset is widely used within the literature and will be used to train our models which count the total number of people present in an image and evaluate their performance in a diverse range of scenarios. This dataset is directly related to some of the MARVEL use cases (Section 2.1, V, VI, and IX) and can be useful in the training of models.

4.1.6 DISCO dataset[27]

The DISCO dataset (Hu et al 2020) contains 1,935 images of 1024×1980 (Full HD) resolution and a 1-second audio clip corresponding to each image, starting 0.5s before the frame was captured and ending 0.5 afterwards. The images are taken from a variety of perspectives, both during day and night, and contain both high-density and low-density populations. DISCO is currently the only publicly available dataset for audio-visual crowd counting. The goal of audio-visual crowd counting is to use the ambient audio stream to obtain a more accurate count, particularly in situations where the image quality is low, for instance, low resolution, low illumination, severe occlusion, and equipment noise. This dataset will be used to train and evaluate our audio-visual crowd counting models. This dataset is directly related to some of the MARVEL use cases (Section 2.1, V, VI, VIII and IX) and can be useful in the training of models.

4.2 MARVEL datasets

This section describes the datasets that are developed as part of the MARVEL project. The project partners (GRN, MT, UNS) will collate, develop and curate these models which will be used for the training or fine-tuning of the models. Details on each dataset are given in the following sections. Parts or all of these datasets will be made publicly available, for use by researchers.

[25] https://svip-lab.github.io/datasets.html

[26] http://www.ee.cuhk.edu.hk/~xgwang/expo.html

[27] https://zenodo.org/record/3828468#.YNAY3RMzZBw

4.2.1 GRN-AV-traffic-entity

This dataset (GRN-Audio-Visual-traffic-entity) is intended for training both unimodal and multimodal signal processing and machine learning models that detect entities making use of public roads and in addition extract the trajectory of the same entities as they move across a junction. These models are expected to output csv structured data that are further processed to implement the respective use cases, namely (I, II, III, IV, VII, VIII) as described in section 2.1. In addition, one example of a post-processing task is anomaly detection over the csv data. The audio-visual data is collected using the GRNEdge setup described in section 3.3.1. The audio-visual recordings take place on public roads, Figure 9. In addition, some of the videos may be sourced from public repositories with the appropriate distribution license.

(a) (b)

Figure 9. Example locations for data collection in Malta, (a) urban junction, and (b) sub-urban junction

The audio track is manually annotated using ELAN (section 3.3.2). The ontology used in the annotation is obtained by extending the MAVD (Zinemanas et al 2019) ontology, which is defined by Entity and Component taxonomies and Actions that link an entity to a component. The sets are the following:

Entities = {Car, Motorcycle, Light Goods vehicle, Heavy goods vehicle, Bicycle, Pedestrian, Micro-mobility, Bus, Pedelec, Other}
Component = {Brakes, Engine, Wheel, Door, Horn, Alarm, Compressor, Bell, Voice, Footsteps, Motor, Music}
Actions = {Rolling, Whining, Screeching, Playing, Sounding, Shouting, Chatter, Closing, Opening, Idling, Accelerating}

The dataset is made up of a collection of short video clips of 3-5 minutes in length, totalling up to approximately 4 hours of video. The high-level schema for each short clip is (metadata, audio schema, video schema). The schema for the metadata is as follows:

[Setup ID], [Location: GPS coordinates, street address, date, time], [Details on camera: model number, lens, focal length, height, look angle] [Microphone: model number, mono/stereo, type, direction], [Video: frame resolution, frame rate, encoding format], [Audio: sampling rate, quantisation, encoding format].

Each 4-5 minute AV clip and annotations are stored as:

[Location_name_number_of_AV_clip].MKV (stores AV clip)
[Location_name_number_of_AV_clip]_audio.csv (stores audio annotations)

[Location_name_number_of_AV_clip]_video.csv (stores video annotations)

Each AV clip requires approximately 500Mbytes of storage (30GB for the dataset).

The owner of the dataset is GRN and the data will be licensed as CC-BY-NC[28]. The AV data in the dataset will be anonymised (more specifically faces, number plates and human voice are all masked) and the metadata or textual annotations contain no personal data.

4.2.2 GRN-AV-traffic-state

This dataset (GRN Audio-Visual-traffic-state) is intended for the training of multimodal ML classification models that output in one forward pass the traffic state on the road and is primarily intended for efficient traffic analytics including anomaly detection, implemented at the edge on limited computing resources. These models output csv structured data given an AV clip of very short duration, typically 3-5 seconds, and can potentially promise an edge solution in the implementation of use case III. The audio-visual data is collected using the GRNEdge setup described in section 3.3.1. The audio-visual recordings take place on public roads. In addition, some of the videos may be sourced from public repositories with the appropriate distribution license.

The AV clips are annotated with a tool that provides for the viewing of video and the selection of all appropriate labels, in a multilabel setting. The AUD annotation tool, iHEARu-PLAY (section 3.2.1), is a likely candidate. The Ontology used in annotation consists of four classes and a number of gradables per class are defined:

Speed {Standstill, Low, Moderate, High}
Collision {Light, Heavy}
Traffic {Sparse, Moderate, Heavy, Jam}
Stationary Obstacle {Service Vehicle, Breakdown, Accident}

The planned dataset is made up of 500 AV clips (approx. 2GB)

The AV clips and annotations are stored as:

[Location_name_number_of_AV_clip].MKV (stores AV clip)
GRN-AV-traffic-state.csv (stores multi-label annotations for all clips in the dataset)

The owner of the dataset is GRN and the data will be licensed as CC-BY-NC[29]. The AV data in the dataset will be anonymised (more specifically faces, number plates and human voice are all masked) and the metadata or textual annotations contain no personal data.

4.2.3 GRN-TXT-traffic-data

This dataset is generated from AI models (mainly entity detection and tracking) that take as input the audio or video streams and output non-binary structured data, which is then processed by AI models upstream in the pipeline and used in the implementation of some use cases (e.g., III, IV). Furthermore, this dataset is also useful for transport engineers and transport policy-makers, who may want to study/design new infrastructure/policy.

The schema for the csv data is:

[28] https://creativecommons.org/licenses/by-nc/4.0/legalcode

[29] https://creativecommons.org/licenses/by-nc/4.0/legalcode

[record ID], [frame ID], [frame_time], [vehicle_ID], [BB_x], [BB_y], [BB_w], [BB_h], [entity_label], [prediction_confidence], [vehicle_flag]

The owner of the dataset is GRN and the data will be licensed as CC-BY-NC[30]. The metadata and textual annotations contain no personal data.

4.2.4 TrentoOutdoor - Real Recording

TrentoOutdoor is a dataset of real-life recordings and includes data acquired by the surveillance cameras currently mounted in four sites. The 5th site is now monitored by analog cameras that have to be replaced by digital ones. Currently, the data velocity is 1 image per second captured by each camera.

Microphones (IFAG microphone arrays) can be mounted nearby the cameras, but speech has to be removed or made inaudible. Data acquisition from the microphones and storage in the data centre are open issues yet to be solved.

Below are the lists of the cameras available in each site with example images. All cameras feature a third digital way for direct data streaming (analog cameras, movable cameras, and cameras without 3rd access have been removed from the list). Monitoring of crowded areas takes place in Piazza Fiera Christmas Market (Table 5) and Piazza Duomo (Table 6) Weekly Market.

Table 5. Lists of the cameras installed at Piazza Fiera

Nr.	Camera	Type	Optic	Brand	Model	Height	Mike
1	fixed	digital	10-40	Basler	BIP2-1600c-dn	8 m	No
2	fixed	digital	10-40	Basler	BIP2-1600c-dn	8 m	No
3	fixed	digital	10-40	Basler	BIP2-1600c-dn	8 m	No
4	fixed	digital	10-40	Basler	BIP2-1600c-dn	8 m	No

[30] https://creativecommons.org/licenses/by-nc/4.0/legalcode

Table 6. Lists of the cameras installed at Piazza Duomo

Nr.	Camera	Tipo	Optic	Brand	Model	Height	Mike
1	fixed	digital		Basler	BIP2-1600c-dn	40 m	No
2	fixed	digital		Basler	BIP2-1600c-dn	40 m	No
3	fixed	digital		Basler	BIP2-1600c-dn	40 m	No

In both sites installing permanent microphones near the cameras is not feasible because access to the cameras is restricted. Since the weekly market as well as the Christmas market occurs in a defined and limited temporal span, audio recordings can be provided by installing temporal microphones (supervised by a human) in the squares.

Criminal/anti-social behaviours experiments will take place at Piazza S. Maria Maggiore (Table 7).

Table 7. Lists of the cameras installed at Piazza. S. Maria Maggiore – Criminal/anti-social behaviour use case

Nr.	Camera	Type	Optic	Brand	Model	Height	Mike
1	fixed	digital	4-12	Basler	BIP2-1600c-dn	4 m	No
2	fixed	digital	7,5-75	Basler	BIP2-1600c-dn	4 m	Yes
3	fixed	digital	4-12	Basler	BIP2-1600c-dn	7 m	Yes

The three cameras are all equipped with extra ethernet and power connection, allowing the installation of microphones nearby. Microphones will be mounted on 2 cameras, one of which will be monitoring a road (with a pedestrian crossing).

The monitoring of parking places use case will take place at Piazzale Ex Zuffo (Table 8), and the analytics use case can take place at the train station, Table 9.

Table 8. Lists of the cameras installed at Piazzale Ex Zuffo – Monitoring of parking places use case

Nr.	Camera	Type	Optic	Brand	Model	Height	Mike
24	Fixed	digital		Basler	BIP-1600c	8 m	Possible
25	Fixed	digital		Basler	BIP-1600c	8 m	Possible
26	Fixed	digital		Basler	BIP-1600c	8 m	Possible
31	Fixed	digital	3-10,5	Axis	P1435-LE	10 m	Not possible
32	Fixed	digital	10-22	Axis	P1435-LE	10 m	Not possible

All cameras feature the possibility of live streaming. Only three cameras can host microphones.

Table 9. Lists of the cameras installed at Train Station-Piazza Dante – Analytics use case

Nr.	Camera	Type	Optic	Brand	Model	Height	Mike
1	fixed	digitale		Basler	BIP-1600c	4 m	No
2	fixed	digitale		Basler	BIP-1600c	4 m	No
3	fixed	digitale		Basler	BIP-1600c	4 m	No
4	fixed	digitale		Basler	BIP-1600c	4 m	No
5	fixed	digitale	4-12	Basler	BIP2-1600c-dn	10 m	Possible
6	fixed	digitale	4-12	Basler	BIP2-1600c-dn	10 m	Possible
7	fixed	digitale	4-12	Basler	BIP2-1600c-dn	10 m	Possible
8	fixed	digitale	4-12	Basler	BIP2-1600c-dn	10 m	Possible
9	fixed	digitale		Axis	P1425-LE Mk	7 m	No
10	fixed	digitale	4-12	Basler	BIP2-1600c-dn	3,5 m	Possible
11	fixed	Digitale	3,5-10	Basler	BIP-1600c	3,5 m	Possible

4.2.5 TrentoOutdoor - Staged Recording

Data will be recorded with three cameras, three IFAG mems arrays, and three large aperture ST mems arrays. Recordings will take place (probably) in a different location with respect to the use case sites. Different light conditions will be considered: daylight, night with artificial lights, dusk. The dataset will simulate dangerous situations and isolated events related to the above

dangerous situations. The staged dataset will be recorded in local devices and processed off-line. Audio-visual signals will be aligned in a semi-automatic way using state-of-the-art tools and empirical strategies (common impulsive events).

4.2.6 UNS drone dataset

The UNS drone dataset is specifically designed for the MARVEL UNS drone experiment and is intended for the Crowd behaviour classification AI task.

The data collection setup consists of 2 cameras (one on the drone and one in the street) and several microphones in order to perform audio localisation. Current setup includes DJI Matrice 600 drone, RPi v2 camera. MEMS microphones from IFAG will be incorporated. Other camera devices are being considered. Raspberry Pi and Intel NUC will support edge processing and data logging.

Data will be recorded within staged recording process and no existing data will be exploited. Recordings will be performed in squares of UNS campus. Some example locations are given in Figure 10. Video and audio data using common formats (MP4, wav) will be recorded. As the plan is to record about 10 hours of data, at least 10 GB will be required for storage for medium quality. Probably, at least 20 GB of data or even more will be produced.

(a) UNS campus square 1 (b) UNS campus square 2

Figure 10. Some illustrations of the places where recordings are planned at UNS

The dataset consists of 120 videos, each 5 minutes long, which represent three classes of data and two positions for recordings (drone and street). AUD's SensMiner will be used for data annotation and the format for metadata is RDF, using standard vocabularies and ontologies.

The owner of recorded data will be the Faculty of Technical Sciences, University of Novi Sad. It will define access control rights, legal and commercial aspects.

UNS dataset will not contain any sensitive data. Data will be obtained within stage recordings and all participants will sign a written consent.

4.2.7 UNS audio-visual emotion dataset

The dataset will be collected for the UNS use case on emotion audio-visual recognition.

There will be approximately 60 sentences per speaker/emotion combination. The content of the sentences is the same throughout emotions. The database will contain around 20 subjects in each of 5 emotions, who will sign written consent before being recorded. One sentence is 2-4s long, and there will be around 3 minutes of video and audio material per subject/emotion combination. All in all, ~15 minutes per subject, and ~300 minutes the whole database. Although the database will be recorded by different phones, sample rate of audio recordings will be all set to 44.1 kHz (most mobile phones support that sample rate and it can be considered

as high quality). All audio files will be converted to mono and prominent noise will be eliminated. Front cameras, which will be used for recordings differ in quality from phone to phone, but they can usually be considered as low quality, but good enough for the intended purpose. All videos will be converted to the same resolution if required by the algorithm to be used.

The owner of the recorded data will be the Faculty of Technical Sciences, University of Novi Sad. It will define access control rights, legal and commercial aspects.

UNS datasets will not contain any sensitive data. Data will be obtained within staged recordings and all participants will sign a written consent.

5 Analysis of Datasets, Streams and KPIs

Sections 2- 4 described the use cases, the AI tasks, the data collection methods, the annotation methods, and the characteristics of each dataset, including both training datasets and datasets generated by the pilots and the MARVEL framework in real-life. This section provides an analysis and cross-examination of the material described, identifying any gaps and suggesting solutions. In addition, this section takes an initial look as to how the MARVEL data management platform will handle and process the variety of data and paves the way for the following tasks in the project.

5.1 Analysis of Datasets

The datasets can be categorised as (a) open datasets that will be used solely for model development, (b) MARVEL datasets that are useful for training the models and are annotated, (c) MARVEL datasets that are not annotated, and (d) MARVEL datasets that are the final output of the processing stages and are useful for further research and development.

Table 10 indicates the type of AI tasks or AI-based components that are potentially useful for each use case. It is noticed that use cases involving the detection of traffic entities and their trajectory (IV and VIII) are only addressed by the SED/SELD models. One solution is to include a vision-based method for these tasks and another solution is to develop a joint audio-visual model. Likewise, use case XI (UNS emotion recognition) does not benefit from the resident AI tasks and requires the development of additional AI tasks. It can however make use of features obtained from the available feature generation tools.

Table 11 tabulates the applicability of the datasets per use case. Whilst some datasets are intended for specific use cases, other datasets may also be relevant for the same use cases. For example, crowd counting datasets may be useful for most of the use cases and some datasets are useful for pre-training models that are then fine-tuned using specific datasets for the respective use case. In this respect, the annotation scheme of some of the datasets may have to be adjusted to match the fine-tuning process.

Table 12 tabulates the datasets against the AI tasks and gives an indication of any gaps. The training of models for some tasks, such as the anonymisation tasks will rely on additional open datasets (section 2.2.4), while some other tasks such as automated audio captioning (applicable to crowd monitoring) is largely missing in the MARVEL datasets, with the exception of GRN-AV-traffic-state, which is strictly a classification dataset.

The annotation ontology for some MARVEL training datasets is still under development, for example, sound event classes, visual entity classes with respective bounding boxes, and event textual descriptions. In addition, the annotation methodology and tools are not all defined and these will be defined at a later stage.

Some of the AI tasks can potentially benefit more from the fusion of both modalities. In particular traffic entity recognition and event description generation.

The size of the annotated datasets is yet to be determined, even though some have estimated the size. However, the size of these datasets may need to be augmented to improve model accuracy.

Table 10. Matching of AI tasks and AI-based components with use cases

		Use cases									
		Safer Roads	Road User Behaviour	Traffic anomaly events	Traffic Entity Trajectory at Network Junctions	Monitoring of Crowded Areas	Detecting Criminal/anti-social behaviour	Monitoring of parking lots	Analysis of specific areas	Drone Experiment	Emotion Recognition
AI Tasks / AI-based components	**Use case index**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**
	Visual Anomaly Detection	X	X	X		X	X	X		X	
	Audio-Visual Anomaly Detection	X	X	X		X	X	X		X	(X)
	Visual Crowd Counting					X	X			X	
	Audio-Visual Crowd Counting					X	X			X	
	Acoustic Scene Classification			X			X			X	X
	Sound Event Detection	X	X		X	X	X	X	X	X	
	Sound-Event and Localisation Detection	X			X	X	X	X	X	X	
	Automated Audio Captioning					X				X	
	Video Anonymisation	X	X	X	X	X	X	X	X	X	(X)
	Audio Anonymisation	X	X	X	X	X	X	X	X	X	(X)

Table 11. Matching of the datasets to the use cases. The datasets are categorised as: (O) open datasets that will be used solely for model development, (T) MARVEL datasets that are useful for training the models and are annotated, (S-AV) MARVEL audio-visual unlabelled dataset, (S-TXT) MARVEL non-binary structured dataset

		Use cases									
		Safer Roads	Road User Behaviour	Traffic anomaly events	Traffic Entity Trajectory at Network Junctions	Monitoring of Crowded Areas	Detecting Criminal/anti-social behaviour	Monitoring of parking lots	Analysis of specific areas	Drone Experiment	Emotion Recognition
Datasets	**Use case index**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**
	MAVD traffic (O)	X						X			
	UCSD Pedestrian (O)	X									
	Street Scene (O)	X	X	X							
	Shanghai Tech (O)	X			X	X			X	X	
	World Expo ’10 (O)	X			X	X			X	X	
	DISCO (O)	X			X	X					
	GRN-AV-traffic-entity* (T)	X	X	X	X			X	X		
	GRN-AV-traffic-state* (T)			X	X				X		
	GRN-TXT-traffic-data* (S-TXT)		X	X	X			X	X		
	TrentoOutdoor-Real* (S-AV)					X	X				
	TrentoOutdoor-Stage* (T)					X	X				
	UNS Drone* (T)					X				X	
	UNS AV emotion* (T)										X

*Dataset is collected and augmented during the project

Table 12. Matching of the datasets to the AI tasks and AI-based components, assuming the annotations are collected. The datasets are categorised as: (O) open datasets that will be used solely for model development, (T) MARVEL datasets that are useful for training the models and are annotated, (S-AV) MARVEL audio-visual unlabelled dataset, (S-TXT) MARVEL non-binary structured dataset

		AI tasks / AI-based components									
		Visual Anomaly Detection	Audio-Visual Anomaly Detection	Visual Crowd Counting	Audio-Visual Crowd Counting	Acoustic Scene Classification	Sound Event Detection	Sound-Event and Localisation Detection	Automated Audio Captioning	Video Anonymisation	Audio Anonymisation
Datasets	MAVD traffic (O)						X				
	UCSD Pedestrian (O)	X									
	Street Scene (O)	X									
	Shanghai Tech (O)	X		X							
	World Expo '10 (O)	X		X							
	DISCO (O)	X			X						
	GRN-AV-traffic-entity (T)						X		(X)		
	GRN-AV-traffic-state (T)	X	X				X		(X)		
	GRN-TXT-traffic-data (S-TXT)								(X)		
	TrentoOutdoor-Real (S-AV)	X	X							X	X
	TrentoOutdoor-Stage (T)	X	X		(X)		X	X		X	X
	UNS Drone (T)	X	X		(X)		X	X		X	X
	UNS AV emotion (T)		X							X	X

5.2 Analysis of Data Streams

In this section, the velocity and volume of the real-life data streams at the respective edge-fog-cloud layers are estimated from the information given in section 3.4. These estimates are useful to plan the MARVEL framework (section 5.4). Table 13 below summarises the estimates, which together with plans and options made available on where and when the data processing takes place (edge/fog/cloud) inform the task of developing an efficient MARVEL framework to support extreme-scale data analytics.

Table 13. Estimates of data rates per device, layer and location, for devices that are deployed in the streaming of data

Location	Device, Data type and format	Data Velocity	Layer	Data rate per device	Number of Devices	Aggregate Data Rate
Trento (MT)	Camera Video/mp4	1fps	Edge/fog	0.5GB/hour	26	13GB/hour
Trento (MT)	Mono-mic Audio/wav	48kSps	Edge/fog	0.55GB /hour	10*	5.5GB/hour
Trento (MT)	Mic-Array Audio/wav	48kSps x 4	Edge/fog	2.2GB/hour	1*	2.2GB/hour
Trento (MT)	Mic-Array Audio/wav	48kSps x 8	Edge/fog	4.4GB/hour	1*	4.4GB/hour
Trento (MT)	Metadata/text labels/csv	UNK (at this time)	Edge/Fog /Cloud	UNK (at this time)	UNK (at this time)	UNK (at this time)
Malta (GRN)	Camera video/ MKV	30fps 1080p	Edge/fog	5.5GB/hour	4 – 8*	22 – 44GB/hour
Malta (GRN)	Camera video/ MKV	25fps 4k	Edge/Fog	12.5GB /hour	2*	25GB/hour
Malta (GRN)	Mono-mic Audio/mp3	48kSps	Edge/Fog	144MB /hour	10*	1.5GB/hour
Malta (GRN)	Metadata/text labels/csv		Edge/Fog /Cloud	1.5GB/day	4 – 8*	6 – 12GB/day

*Estimated number of devices

5.3 Diversity of data resources

In the MARVEL project, objective I and pillar I address the deployment of novel tools and engineering paradigms for capturing large volumes of real audio-visual data from distributed IoT sources in a smart city environment. In this respect, it is expected that a diversity of IoT devices are connected to the framework and the target of KPI-O1-E1-1, “Different kind of resources to be discoverable”, is to have at least three or more different and diverse resources or IoT devices. In Section 3, microphones, cameras, drones and crowed-sourced mobile applications are specified and some use cases will also make use of weather and accident data. Therefore, it can be concluded that this KPI will be met with high certainty. This however should not stop the framework from accepting or discovering other kinds of data sources, for

example, data collected and processed from social media, from connected vehicles (V2I), or car detection electromagnetic sensors (Section 2.1.8). The latter can be facilitated through a functionality of the data management toolkit (DMT) where the user attaches a new device and the DMT automatically discovers and accepts this device as a new data capturing resource and starts ingesting the data from it. Invariably this would require the DMT to be extendable to a number of device types that are extra to the ones installed during the MARVEL project.

5.4 Initial Plans for Marvel Data Management Platform

The main function of the MARVEL data management platform is to **transfer** and **process** data collected by the data sources or devices in the field across the three system tiers (edge, fog, cloud). Processing at edge/fog layers will allow decision-making output to be conveyed to the end-users without involving the cloud layer. This can take place by deploying AI inference models on edge/fog devices from the cloud, or using Federated Learning techniques, where training is done collaboratively from multiple agents, each contributing a model improvement based on local model/data. From Section 4, two categories of data are identified: Non-binary structured data, e.g., in comma-separated values format, and binary AV (Audio/Video) data.

The MARVEL data management platform will be realised through the tight co-existence of four technological/algorithmic modules, each of which offers certain distinct features, as follows:

DatAna (ATOS Data Acquisition Framework), which has Apache NiFi at its core, provides data ingestion, transformation and interconnection mechanisms for collecting and processing data of different nature and from different sources following a data-flow paradigm. NiFi and MiNiFi offer the possibility of creating topologies of different deployments to connect and move data flows, for instance from the fog to the cloud. NiFi provides out-of-the-box processors for many data sources as well as for the majority of the most common storage and messaging systems, such as SQL and NoSQL databases, Kafka, and others.

StreamHandler (INTRA), which is a high-performance (low latency and high throughput) distributed streaming platform for handling real-time data based on Apache Kafka, is capable of scaling out and accommodating various data from different domains, interoperating with all modern data storage technologies as well as other persistence approaches.

DFB (ITML Data Fusion Bus), which is a data fusion platform for multiple modality data streams, is also based on Apache Kafka, which implements a trustworthy way of transferring large amounts of heterogeneous data between several connected components, and the permanent storage, data aggregation, and analytics.

HDD (CNR Hierarchical Data Distribution) module, decides at run-time the extent of data distribution, depending on the application requirements and constraints, through smart and adaptive algorithmic techniques.

Depending on the specific infrastructure deployed for the experiments, both centralised and distributed deployments will be considered; in the distributed case, the edge/fog nodes will locally share some burden of the computations, which increases scalability at the expense of taking possibly myopic decisions due to the lack of a global vision of the system.

It is thought that Apache Kafka will be a reference technology within the MARVEL data management platform since two of the modules are natively based on that platform (StreamHandler and DFB), while DatAna uses Apache NiFi, which already provides connectors and integration tools with Apache Kafka. It may also be possible to cooperate with Apache Flink to scale the data processing and ML aspects in the cloud connecting to (e.g.) Apache Spark Streaming. HDD is expected to design and implement the connectors and the data

management algorithms with and among the various modules. The co-existence and the complementarity of the technological modules will be performed in T2.2 which has just started.

Based on current technologies, the system can definitely handle non-binary data streams, including events, metadata, structured or unstructured documents, logging and other information that stream into the platform. Within the context of MARVEL, any component that produces non-binary data can be directly connected to and stream their data into the system. These components can either collect non-binary data directly from edge devices, or middleware, data processing components that stream their output to the system. An example for the former case would be the collection of log files, performance metrics, metadata or other non-binary information from edge devices. An example for the latter case would be the connection with a First- or Second-Layer AI component. These components may perform any suitable ML algorithm on either non-binary or AV data and produce outputs such as classification, clustering results, or any type of data that may be used for predictions or decision-making. The system will be able to handle the estimated volume generated for non-binary data (Section 5.2 – Table 13), assuming 10s or a few 100s of devices, depending on the velocity of the devices, which is stochastic in nature.

On the other hand, a technology gap has been identified in the handling of AV data, which has some unique characteristics that make them different from structured non-binary data, which are the typical sources in Apache Nifi and Apache Kafka, laying at the core of the data management platform. In particular:

- AV data streams are continuous: breaking them into chunks of data, which may be necessary to fit for technical compatibility reasons, creates artificial discontinuities that might have an impact on performance. For instance, in case of a video stream, a given event of interest might take place in between two chunks: the data analytics system might detect two different events, instead of one, or miss it completely due to insufficient data to trigger inference in any.
- AV data streams have a much higher volume and velocity than other typical sensing applications, which might violate some of the underlying design assumptions of the target platforms identified.
- AV data streams might need to be synchronised in time and space to provide a meaningful decision output. For instance, it can take place only if the combined video and audio in a scene can trigger a given alert, whereas any of them alone would not. The issue of synchronisation is particularly relevant if processing occurs in the edge/fog layers: due to the hardware limitations of devices running the computation, processing of a number of related sources might have to be distributed across multiple nodes, which would make it very difficult to jointly process them.

As mentioned, the data platforms based on Apache NiFi and Apache Kafka are better suited to handle non-binary data. Handling video and audio in edge/fog devices could potentially be done by using native graphical processing capabilities of the fog nodes, more suited to the task.

- For video processing: For instance, a NVIDIA Jetson device (yet to be selected) or similar device can capture the video and audio and run it through live DL apps or analytics pipelines at the fog layer (i.e., anonymisation, classification, synchronisation, etc.). Then the results of these processes can be written to a JSON or CSV file and the Apache NiFi or MiNiFi on the fog layer could tail the file to stream the classifications to another NiFi central node, from where it can be streamed via Kafka to other data management platforms, Flink, etc.
- Moving large files to the cloud: Apache NiFi might move large binary files from one NiFi (or MiNiFi) node to another using the NiFi site-to-site protocol (S2S). This is a potential choice, although as NiFi is based on Java, either the Java Heap size (to be estimated during

project) should be big enough to handle the size of the binary files processed simultaneously, or the files should be split into several smaller files to be transferred and then recomposed. This could be an option if the simultaneous number of video files to transfer is minimal or the size of the files is not huge. This would require testing and a reasonably sized NiFi cluster in the cloud to handle large data coming from multiple devices equipped with MiNiFi or NiFi at the fog layer.

6 Guidelines for privacy assurance and anonymisation

To ensure that the collection of data is compliant with the GDPR regulations and all privacy, ethical and legal concerns are adequately addressed, the project will make use of tools and algorithms to anonymise the data where and when necessary. This section provides guidelines on the anonymisation for the MARVEL training datasets as well as for live streaming of data, where the intention is to execute algorithms on low-powered resources. The latter is one of the main challenges in the MARVEL project.

6.1 Privacy assurance and anonymisation in use cases

All use cases and pilot executions (GRN, MT, UNS) involve audio and video recordings of real-life events in public spaces, namely spaces where people are known to gather, such as town squares, and road and transport physical infrastructure, such as road network junctions, streets and parking lots. The audio-video recording devices are typically mounted on stationary poles and buildings, with the exception of the drone experiment, which will make use of both drone-mounted cameras and microphones as well as devices on the ground, either fixed to physical infrastructure or even mobile.

The devices will be recording real-life and invariably will also record, PD, which data is not required to execute the pilots. To provide privacy assurance and adherence to GDPR regulations, all the audio and video recordings will be anonymised (off-line and in real-time) in terms of masking/obfuscation of human faces, vehicle number plates, and human voice, using the techniques developed in line with the description in section 6.2. For the case of the data collection process that will involve actors, such as the drone experiment (UNS) and staged data collection experiments in the City of Trento (MT), all participants in the experiment will sign a written consent and anonymisation techniques may not be required for these specific cases. In addition, it is the intention of the MARVEL project to carry out anonymisation at the edge, thus simplifying system complexity and reducing the risk of personal data breach.

6.2 Anonymisation Tools and algorithms

This section describes how the project intends to address PD for both the data at rest (off-line anonymisation of the MARVEL training datasets) and data in motion (real-time anonymisation during the live streaming of data)

6.2.1 Off-line Anonymisation Tools and algorithms

Video anonymisation is mostly achieved by obfuscating within any video stream the Personal Data (PD) which can be used to derive the person's identity. Faces are the most concerning visual data for persons while vehicle number plates for vehicles. To perform video redaction, object detection/segmentation often serves as the first step, where the YOLO series, e.g., scaled yolov4 (Bochkovskiy et al, 2020), are the most popular object detectors in practical applications for its real-time and accurate performance. GRN has experimented with such a face detection and masking system which performed as expected on off-line data. Different from object detectors that often output the detections in the format of bounding boxes, instance segmentation aims to segment out the exact contours of the objects, which could be ideal for obfuscating the personally identifiable data while leaving the contextual information as much as possible intact. Recent methods, such as SipMask++ (Cao et al, 2020), YOLACT++ (Bolya et al, 2020) have greatly improved the real-time performance, but at a cost of dropped segmentation accuracy compared to the current SOTA accuracy, leaving it a challenging task to run on resource-limited machines.

Once objects, e.g., faces or car plates, are detected, obfuscation methods can be applied to anonymise the visual content. The most common irreversible options include image blurring via filters (Du et al, 2019), pixelation by enlarging the pixels (Gerstner et al, 2016), mosaic by merging small blocks of pixels from different regions, masking by simply replacing the visual information with other data. Moreover, visual abstraction techniques replace the persons appearing in images with a visual model (Padilla-lopez et al, 2015), with the advantages of not only preserving privacy but also enabling further analysis on human activities. MARVEL will further advance and develop the abstraction techniques based on GAN-based approaches by converting the face of one identity to another non-sensitive identity to preserve the image content though at a higher computational cost. The data from staged recordings will serve as a key enabler of the development of the algorithm for training and testing.

Audio anonymisation aims to remove any information about the speaker identity from an audio stream. In practice, voice anonymisation begins with detecting audio segments that contain voice. To this end, AUD's voice activity detection technology (VAD) can be used. The technology is built on the openSMILE toolkit and can be used both in streaming and batch fashion. For convenience, a python package has been built around the batched version and has now been used by GRN for initial off-line processing. In the initial phase of the project, segments containing voice can be completely removed in order to avoid potential privacy violations. The quantity of the remaining data should be sufficient to collect enough data for training.

When audio streams are being processed on a remote server for some services, one popular strategy that is also bandwidth efficient consists of extracting features on the sensing device and removing from them any information about the speaker's identity. This can be achieved using neural feature extractors adversarially trained on the speaker identity, as done in (Srivastava et al, 2019) for speech recognition. One of the limitations of this approach is that different feature sets may be required by the processing services, as for example audio captioning, sound event detection, speech activity detection, etc. This issue can be addressed using general purpose features (LEAF (Zeghidour et al, 2021), PASE+ (Ravanelli et al, 2020)) adversarially trained.

Another strategy aims at removing speaker identity information from the audio stream while preserving all the other information about content and context. One approach recently proposed at the Voice Privacy Challenge implements a formant shifting strategy based on McAdams coefficients which is effective and computationally light (Patino et al, 2020). Most successful systems implement a cascade of speech-to-text, xvector transformation (or other speaker related features) and text-to-speech (Fang et al, 2019). The latter has been recently implemented in a seq2seq fashion (Huang et al, 2020). These approaches suffer from recognition errors. A similar anonymisation result can be obtained by applying voice conversion (Yi et al, 2020) whose goal is to clone a speaker voice into another speaker, similarly to visual abstraction. Voice conversion methods have been proposed with starGANs (Kameoka et al, 2018), and conditional VAE (Kameoka et al, 2019), just to mention a few. These methods typically need training for both source and target speakers (not parallel material where speakers utter the same sentence). Both anonymisation and conversion strategies have only been applied to close-talking signals, and are not robust enough in real outdoor environments.

In the future, the process can be improved to make more data available. Instead of removing segments containing voice, the segments can be filtered either to completely remove voice components, or to delexicalize them. Removing voice components is equivalent to "inverse" speech enhancement. Several open-source enhancement algorithms already exist and can be used out-of-the-box. After running them on an input audio file, the output can be subtracted

from the original audio to remove, instead of preserve, voice. Moreover, new algorithms can be developed targeting the removal of voice in particular. Alternatively, identifying information can be removed by voice conversion, i.e., changing the voice of the speaker using appropriate algorithms, as reviewed above. However, voice conversion does not remove the spoken content of the utterance, which may potentially contain identifying information (e.g., names/addresses/etc.) and as such may not be the appropriate technique.

6.2.2 On-line Anonymisation Algorithms

Preserving the privacy of individuals in real-time streaming can be carried out by either first anonymising the audio-video data as described in Section 6.2.1 and then processing the anonymised data, or by processing the non-anonymised data at the first processing stage (e.g., edge) guaranteeing an output that does not contain any personal data.

Following the *privacy by design* (PbD) paradigm (Schaar, 2010), adopted in the MARVEL project, data can be anonymised by extracting useful, non-sensitive information that can be safely shared. However, anonymising video and audio streams in real-time is a computationally-intensive process, and state-of-the-art techniques rarely fit on the typically low-resource devices that constitute the edge. It is thus a significant challenge to strike the right balance between privacy, and efficient engineering design. A modern Edge-to-Fog-to-Cloud (E2F2C) architecture, such as the one envisioned by MARVEL, should intelligently allocate resources to optimally balance this trade-off between privacy and computational complexity. Edge nodes should dynamically decide how much of the anonymisation measures implemented by MARVEL they should perform themselves (if any), and delegate the rest to an upstream node, for example in the fog layer.

Another way to follow the PbD principle is to implement data processing on edge devices. In fact, by exploiting processing pipelines in which the algorithm's output is in line with the privacy regulations and the input is not stored or transmitted, we can achieve a privacy-aware perception system. There are already many implementations of edge processing in literature, for both video, and audio signals. For video, the input is analysed frame by frame without the need to store sensitive data. Depending on the final application of the system, the edge device can transmit the trajectories of the objects in the scene (Paissan et al, 2021), and embeddings which are afterwards used to extract the needed information or higher-level information (Cerutti, Milosevic, & Farella, 2018).

For audio, there is the need to process within more sensitive time frames using sliding windows since the window length impacts the pipeline accuracy. For generic applications (e.g., sound-event detection), the window length (around 1 second) does not imply the need for data storage. Therefore, the same approach used for video can be exploited to share only pre-processed, privacy-preserving data (Cerutti et al, 2020; Cerutti, 2020b). As mentioned in section 6.2.1, AUD's VAD technology can already work in streaming fashion, and thus can be already used in an edge or fog node.

It is clear that the solution is dependent on both the characteristics and needs of the use case and on the computational efficiency of the anonymisation and event detection algorithms. For example, in some use cases it is beneficial to preserve the (anonymised) audio-video clip depicting the event and make it available to human operators in traffic or law-enforcement control centres, with the possibility of pointing to the original audio-video data in the latter case. On the other hand, in other use cases there is no need to keep the raw data for storage since privacy-preserving features suffice (e.g., motion vectors or a list of time-stamped events). This eases the burden on anonymisation algorithms, as the features will not contain protected information that needs to be removed. This alternative is only viable if the models built in WP3

and WP4 depend on such features. On the other hand, training dataset augmentation invariably requires the preservation of the anonymised raw audio-video data. The solution, therefore, requires the distribution of both anonymisation and event detection algorithms executing along the edge, fog, and cloud layers, such that privacy-preserving features make up the bulk of the data in the cloud.

7 Conclusions

This document dealt with the collection and analysis of the experimental datasets that MARVEL will generate to foster research and development of new data-driven disruptive applications in smart cities. The datasets are motivated by first describing use cases (pilots that will demonstrate the use of the MARVEL data management platform) and the AI tasks that are required to implement the use cases. It is to be noted that the AI tasks are reusable for other applications and use cases. The data collection methodology and, where applicable, the annotation schemes are described, followed by the details of each dataset, that include the relevant technical characteristics as well as ownership of data, accessibility, etc. Invariably, data collecting devices deployed in smart cities pick-up personal data which has to be anonymised in order to guarantee a data sharing process that ensures personal privacy and complies with GDPR legislation. This required the identification of PD in the dataset, and various audio and video anonymisation methods are put forward and discussed for implementation. Finally, an analysis of the components described is carried out, including a look into how the MARVEL data management platform will deal with the variety of data.

The annotation ontology for some MARVEL training datasets is still under development, related to for example the sound event categories and visual object categories. In addition, the annotation methodology and tools are not all identified, for example in the case of sound textual descriptions and object bounding boxes. Some of the AI tasks can potentially benefit more from the fusion of both modalities. In particular, traffic entity recognition and event description generation. The size of the annotated datasets is yet to be determined, even though estimates are provided. However, the size of these datasets may need to be augmented (T2.3 in WP2) to improve the performance of the models. With regards to the data management platform, a technology gap in the handling of AV data has been identified. AV data has some unique characteristics that make them different from structured non-binary data, which are the native data sources handled in Apache Nifi and Apache Kafka. The document, therefore, proposes some solutions for the handling of AV data.

8 References

Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y. M. (2020) "YOLOv4: Optimal speed and accuracy of object detection," arXiv [cs.CV]. Available at: http://arxiv.org/abs/2004.10934.

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). YOLACT++: Better real-time instance segmentation. arXiv preprint arXiv:1912.06218.

Cai, Y. (2003, June). "How many pixels do we need to see things?". International Conference on Computational Science (pp. 1064-1073). Springer, Berlin, Heidelberg.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE transactions on pattern analysis and machine intelligence, 43(1), 172-186.

Cao, J., Anwer, R. M., Cholakkal, H., Khan, F. S., Pang, Y., & Shao, L. (2020). SipMask: Spatial Information Preservation for Fast Image and Video Instance Segmentation. arXiv preprint arXiv:2007.14772.

Cerutti, G., Milosevic, B., & Farella, E. (2018, June). Outdoor people detection in low resolution thermal images. In 2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech) (pp. 1-6). IEEE.

Cerutti, G., Prasad, R., Brutti, A., & Farella, E. (2020). Compact recurrent neural networks for acoustic event detection on low-energy low-complexity platforms. IEEE Journal of Selected Topics in Signal Processing, 14(4), 654-664.

Cerutti, G., Andri, R., Cavigelli, L., Farella, E., Magno, M., & Benini, L. (2020, August). Sound event detection with binary neural networks on tightly power-constrained IoT devices. In Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design (pp. 19-24).

Du, L., Zhang, W., Fu, H., Ren, W., & Zhang, X. (2019). An efficient privacy protection scheme for data security in video surveillance. Journal of Visual Communication and Image Representation, 59, 347-362.

Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N., & Bonastre, J. F. (2019). Speaker anonymization using x-vector and neural waveform models. arXiv preprint arXiv:1905.13561.

Gerstner, T., DeCarlo, D., Alexa, M., Finkelstein, A., Gingold, Y., & Nealen, A. (2013). Pixelated image abstraction with integrated user constraints. Computers & Graphics, 37(5), 333-347.

Hu, D., Mou, L., Wang, Q., Gao, J., Hua, Y., Dou, D., Xiang Zhu, X. (2020). "Ambient Sound Helps: Audiovisual Crowd Counting in Extreme Conditions", arXiv:2005.07097.

Huang, W. C., Hayashi, T., Watanabe, S., & Toda, T. (2020). The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading asr and tts. arXiv preprint arXiv:2010.02434.

Jack, R., Garrod, O., and Schyns, P. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. Curr. Biol. 24, 187–192. doi: 10.1016/j.cub.2013.11.064.

Jovičić, S.T., Kašić, Z.: Đorđević, M., Rajković M.: Serbian emotional speech database: design, processing and evaluation. In: Proc. of the SPECOM 2004: 9th Conference Speech and Computer, St. Petersburg, Russia, pp. 77–81 (2004)

Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018, December). Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 266-273). IEEE.

Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2019). ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(9), 1432-1443.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). "Deep learning face attributes in the wild". Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). "Anomaly detection in crowded scenes". IEEE Conference on Computer Vision and Pattern Recognition.

Padilla-López, J. R., Chaaraoui, A. A., & Flórez-Revuelta, F. (2015). Visual privacy protection methods: A survey. Expert Systems with Applications, 42(9), 4177-4195.

Paissan, F., Gottardi, M., & Farella, E. (2021). Enabling energy efficient machine learning on a Ultra-Low-Power vision sensor for IoT.

Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., & Evans, N. (2020). Speaker anonymisation using the McAdams coefficient. arXiv preprint arXiv:2011.01130.

Ramachandra, B. and Jones, M.J. (2020). "Street Scene: A new dataset and evaluation protocol for video anomaly detection". IEEE Winter Conference on Applications of Computer Vision.

Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., & Bengio, Y. (2020, May). Multi-task self-supervised learning for robust speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6989-6993). IEEE.

Rehman, A. U., Ullah, H. S., Farooq, H., Khan, M. S., Mahmood, T. and Khan, H. O. A. (2021). "Multi-Modal Anomaly Detection by Using Audio and Visual Cues". IEEE Access, vol. 9, pp. 30587-30603.

Salamon, J., MacConnell, Cartwright, D. M., Li, P. and Bello, J. P. (2017). "Scaper: A library for soundscape synthesis and augmentation," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).

Schaar, P. (2010). Privacy by design. Identity in the Information Society, 3(2), 267-274

Srivastava, B. M. L., Bellet, A., Tommasi, M., & Vincent, E. (2019). Privacy-preserving adversarial representation learning in ASR: Reality or illusion?. arXiv preprint arXiv:1911.04913.

Yi, Z., Huang, W. C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., ... & Toda, T. (2020). Voice Conversion Challenge 2020—Intra-lingual semi-parallel and cross-lingual voice conversion—. In Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020 (pp. 80-98).

Zeghidour, N., Teboul, O., Quitry, F. D. C., & Tagliasacchi, M. (2021). LEAF: A Learnable Frontend for Audio Classification. arXiv preprint arXiv:2101.08596.

Zhang, C., Li, H., Wang, X., and Yang, X. (2015). “Cross-scene crowd counting via deep convolutional neural networks”. IEEE Conference on Computer Vision and Pattern Recognition.

Zhang, Y., Zhou, D., Chen, S., Gao, S. and Ma, Y. (2016). “Single-Image Crowd Counting via Multi-Column Convolutional Neural Network”. IEEE Conference on Computer Vision and Pattern Recognition.

Zinemanas, P., Cancela, P. and Rocamora, M. (2019). "MAVD: a dataset for sound event detection in urban environments". DCASE 2019 Workshop, 25-26, New York, USA http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop_Zinemanas_70.pdf