|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding by Artificial Intelligencewww.mpai.community |

|  |  |
| --- | --- |
| **N697** | 2022/05/18 |
| **Source** | Requirements (CAV) |
| **Title** | MPAI-CAV Environment Sensing Subsystem (ESS) |
| **Target** | MPAI-20 |

Contents

[1 Use Case description 1](#_Toc103775018)

[2 Reference architecture 2](#_Toc103775019)

[3 Input and output data 2](#_Toc103775020)

[4 AI Modules 3](#_Toc103775021)

[5 I/O Data summary 3](#_Toc103775022)

[6 Data formats 4](#_Toc103775023)

[6.1 Audio Objects 4](#_Toc103775024)

[6.2 Basic World Representation 4](#_Toc103775025)

[6.2.1 Introduction 4](#_Toc103775026)

[6.2.2 The CAV’s Environment Sensing Technologies 5](#_Toc103775027)

[6.2.3 Requirements for WR integration 5](#_Toc103775028)

[6.2.4 Creation of a WR 5](#_Toc103775029)

[6.2.5 EST formats 6](#_Toc103775030)

[6.2.6 WR integration requirements 6](#_Toc103775031)

[6.2.7 FWR requirements 7](#_Toc103775032)

[6.3 GNSS Coordinates 8](#_Toc103775033)

[6.4 GNSS Data 8](#_Toc103775034)

[6.5 Lidar Data 8](#_Toc103775035)

[6.6 Microphone Array Audio 9](#_Toc103775036)

[6.7 Moving Object Tracker Data 9](#_Toc103775037)

[6.8 Offline Maps 9](#_Toc103775038)

[6.9 Radar Data 9](#_Toc103775039)

[6.10 State 10](#_Toc103775040)

[6.11 Traffic Signalisation 10](#_Toc103775041)

[6.12 Ultrasound Data 11](#_Toc103775042)

[6.13 Ultrasound Objects and Scene (Ultrasound) 11](#_Toc103775043)

[6.14 Video Camera data 11](#_Toc103775044)

[6.15 Visual Objects and Scene (Camera) 12](#_Toc103775045)

[6.16 Visual Objects and Scene (Lidar) 12](#_Toc103775046)

[6.17 Visual Objects and Scene (Radar) 12](#_Toc103775047)

[7 Terminology 12](#_Toc103775048)

[Annex 13](#_Toc103775049)

# Use Case description

The purpose of the ESS is to acquire all sorts of electromagnetic and acoustic data directly from its sensors and other physical data of the Environment (e.g., temperature, pressure, humidity etc.) and of the CAV (Pose, Velocity, Acceleration) from Motion Actuation Subsystem with the main goal of creating the Basic World Representation.

# Reference architecture

*Figure 1* gives the Environment Sensing Subsystem Reference Model.



*Figure 1 – Environment Sensing Subsystem Reference Model*

The typical series of operations carried out by the Environment Sensing Subsystem (ESS) is given below. The sequential description of steps does not imply that an action is only carried out after the preceding one has been completed.

1. The CAV gets its Pose and other environment data from:
	1. Global Navigation Satellite System (GNSS).
	2. Vehicle Localiser in ESS.
	3. Other Environment data (e.g., weather, air pressure etc.).
	4. Coordinates data (Pose, Orientation and their time derivatives).
2. The CAV creates a Basic World Representation (BWR) by:
	1. Acquiring available Offline maps of its current Pose.
	2. Fusing Visual, Lidar, Radar and Ultrasound data.
	3. Updating the Offline maps with
		1. Other static objects.
		2. All moving objects.
		3. All traffic signalisations.
3. The CAV compresses and stores a subset of the sensor data on board the ESS.

# Input and output data

*Table 1* gives the input/output data of Environment Sensing Subsystem.

*Table 1 – I/O data of* *Environment Sensing Subsystem*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Pose-Velocity-Acceleration | Motion Actuation Subsystem  | To be fused with GNSS data |
| Other Environment Data | Motion Actuation Subsystem  | Temperature etc. to be added to Basic World Representation |
| Global Navigation Satellite System (GNSS) | ~1 & 1.5 GHz Radio | Get Pose from GNSS  |
| Radio Detection and Ranging (RADAR) | ~25 & 75 GHz Radio | Get RADAR view of Environment |
| Light Detection and Ranging (LIDAR) | ~200 THz infrared | Get LiDAR view of Environment |
| Ultrasound | Audio (>20 kHz) | Get 20 kHz view of Environment |
| Cameras (2/D and 3D) | Video (400-800 THz) | Get visible view of Environment |
| Microphones | Audio (16 Hz-16 kHz) | Get Audible view of Environment |
| **Output data** | **To** | **Comment** |
| State | Autonomous Motion Subsystem | For Route, Path and Trajectory |
| Basic World Representation | Autonomous Motion Subsystem | Locate CAV in Environment |

# AI Modules

*Table 2* gives the AI Modules of Environment Sensing Subsystem.

*Table 2 – AI Modules of* *Environment Sensing Subsystem*

|  |  |
| --- | --- |
| **AIM** | **Function** |
| **GNSS Data Coordinate Extractor** | Computes global coordinates of CAV. |
| **Radar Data Processor** | Extracts visual scene and objects from radar. |
| **Lidar Data Processor** | Extracts visual scene and objects from lidar. |
| **Ultrasound Data Processor** | Extracts visual scene and objects from ultrasound. |
| **Camera Data Processor** | Extracts visual scene and objects from camera. |
| **Environment Sound Data Processor** | Extracts audio scene and objects from microphone. |
| **Environment Data Recorder** | Compresses/records a subset of data produced by CAV sensors and processed data at a given time. |
| **Vehicle Localiser** | Estimates the current CAV State in the Offline Maps. |
| **Moving Objects Tracker** | Detects, tracks and represents position and velocity of Environment moving objects. |
| **Traffic Signalisation Recogniser** | Detects and recognises traffic signs to enable the CAV to correctly move in conformance with traffic rules. |
| **Basic World Representation Fusion** | Creates Basic World-Representation by fusing Offline Map, moving and traffic objects, and other sensor data |

# I/O Data summary

For each AIM (1st column), *Table 3* gives the input (2nd column) and the output data (3rd column). The following 3-digit subsections give the requirements of the data formats in columns 2 and 3.

*Table 3 – Environment Sensing Subsystem data*

|  |  |  |
| --- | --- | --- |
| **AIM or Subsystem** | **Input** | **Output** |
| **Vehicle Localiser** | GNSS Coordinates | State |
| Pose-Velocity-Acceleration |
| Offline Maps |
| **Environment Recorder** | State | -- |
| Basic World Representation |
| Other Environment Data |
| **GNSS Coordinate Data Extractor** | GNSS data | Global coordinates |
| **Radar Data Processor** | Radar data | Visual Objects and Scene |
| **Lidar Data Processor** | Lidar data | Visual Objects and Scene |
| **Ultrasound Data Processor** | Ultrasound data | Visual Objects and Scene |
| **Camera Data Processor** | Camera data | Visual Objects and Scene |
| **Microphone Sound Data Processor** | Microphone data | Sound Objects and Scene |
| **Traffic Signalisation Detector** | Visual Objects and Scene | Traffic signalsTraffic rules |
| **Moving Objects Tracker** | Visual Objects and Scene | Moving objects’ states |
| **Basic World Representation Fusion** | State | Basic World Representation |
| Offline maps |
| Visual Objects and Scene |
| Audio Objects and Scene |
| Environmental Data |
| Static and moving objects |
| Traffic signals |

# Data formats

## Audio Objects

The sound field of the Environment is captured by the external Microphone Array. Objects are extracted and added to the Basic and eventually Full World Representation after receiving information from other CAVs in range.

**To Respondents**

Respondents are requested to propose an Audio Objects and Scene Format that provides information about audio objects identified in the Environment with their semantics and the degree of accuracy with which Audio Objects have been represented.

## Basic World Representation

### Introduction

The World Representation (WR) is the basic data format enabling a CAV to achieve its mission. It is defined as “A digital representation of the Environment produced by an Environment Sensing Technology (EST) of a CAV or an Offline Map or an integration thereof”. A WR may be a rather complex data structure that includes several elementary components called Data Types each having a Data Format.

The CAV’s WR, called Basic World Representation (BWR) results from the *integration* of the different WRs generated by different ESTs of a CAV. A CAV may produce its Full World Representation (FWR) by *integrating* the BWRs received from other CAVs in range and produce.

### The CAV’s Environment Sensing Technologies

The ESTs considered by MPAI-CAV are:

1. Global navigation satellite system or GNSS (~1 & 1.5 GHz Radio).
2. Geographical position and orientation, and their time derivatives up to 2nd order.
3. Data in the visible range, possibly supplemented by depth information (400 to 700 THz).
4. Lidar (~200 THz infrared).
5. Radar (~25 & 75 GHz).
6. Ultrasound (> 20 kHz)
7. Sound in the audible range (16 Hz to 16 kHz).
8. Other environmental data (temperature, humidity, ...)

### Requirements for WR integration

This document targets the definition of a standard WR, with the understanding that it should be possible to losslessly convert an EST-specific WR to the *standard* WR. This does not mean that each EST shall use the standard format but that the EST uses a compatible format. WR compatibility is defined as “The ability of a WR to be converted to another WR without loss of information”.

WR compatibility has the following features:

1. WR1 may include Data Types not included in WR#2 and vice versa.
2. The Data Types in WR#1 need not have the same values as those in WR#2, i.e., their values may:
	1. Not have the same Accuracy.
	2. Conflict even though their Data Types may be the same or compatible.

WR integration shall satisfy the following requirements:

1. A CAV may use a set of ESTs each providing a WR each having its own format. However, the format of the WRs shall be the same, or compatible.
2. Different ESTs may provide the same or different parameters with different levels of Accuracy.
3. Two CAVs in range may use two different EST sets. However, their BWRs shall have the same or compatible formats.
4. The BWRs of two CAVs with different positions and orientations will be different and have different levels of detail or have conflicting values, but the two BWRs shall have the same or compatible formats.
5. The formats of the digital maps shall allow for integration in a WR without loss of information.

### Creation of a WR

For the purpose of the following analysis, it is assumed that an EST outputs a WR that is

1. Time-dependent and constantly updated object-based.
2. All components of the sensed Environment are objects.
3. Object have different resolutions, e.g., the background is a single object with low resolution while a CAV 5 m from the ego CAV is given with high resolution.

The WR of an EST shall have one of a limited number of formats standardised by MPAI, e.g.,

1. Object coordinates.
2. Bounding box.
3. 2D object description.
4. 3D object description, i.e., Volumetric.

An EST assigns the following attributes to each object sensed from the Environment:

1. Identifier.
2. Timestamp.
3. State (Static, Dynamic, Unknown).
4. Position attributes (position, velocity, and acceleration).
5. Orientation attributes (orientation, velocity, and acceleration).
6. Identifiers of objects existing at previous times and generating the current object.
7. Format.
8. Semantics.

Attributes of position, orientation, and objects identifiers (6.) are given with an estimated accuracy.

The process whereby an EST creates WR(t) from sensed data SD(t) and WR(t-Δt) is:

1. Computes the prediction of WR(t) using WR(t-Δt), WR(t-2Δt) etc.
2. Compares each predicted object in the predicted WR(t) with each object in the sensed data.
3. Creates WR(t) by
	1. Updating the attributes of each object inherited from previous WRs.
	2. Removing objects present in previous WRs and no longer present in WR(t).
	3. Adding and assigning attributes to new objects sensed at time t (entirely new object of the merge of two objects).

Each EST has its own specific memory of the past WRs. Objects in the far past may have coarser attributes.

The EST WR format is a list of objects with their attributes.

The BWR results from the merge of the individual EST WRs.

To help each EST to provide a high-quality WR, a subset of the BWR can be fed back to each EST.

### Road topology creation

Some ESTs (e.g., cameras) allow for the identification of road signs. They can produce a road topology that is separate from the EST WR.

### External object motion tracking

In some case and EST will only provide snapshots containing objects without object motion tracking. stark@eecs.umich.edu Wayne Stark Uhnder

### EST formats

The format used by an EST can be 2D, 2.5D or 3D.

**2D World Representation**:

1. Static environment:
	1. Parametric free space representation represented as a single object.
	2. Alternative representations as individual static objects.
2. Dynamic environment: object-based representation.

**2.5D World Representation**:

1. Static components of the scene
	1. Grid-based (elevation maps or stixel world), represented as a single object.
	2. Object-based for traffic poles and signals (e.g., Stixel world, Multi-level surface map).
2. Object-based for the dynamic parts (e.g., Stixel world, Multi-level surface map).

The **3D World Representation** requirements are:

1. Static components of the scene
	1. Voxel grids, meshes, possibly as a single.
	2. Object-based for traffic poles and signals (voxel grids, meshes).
2. Dynamic components of the scene (point clouds, voxel grids, meshes,…)

### WR integration requirements

Figure 1 represents the integration of different environment data provided by different ESTs of a CAV into a BWR.



Figure 1 – Sensed WRs from different ESTs and the Basic World Representation

The BWR requirements are:

1. The BWR shall have a format standardised by MPAI.
2. The BWR shall use:
	1. All sensed Audio and Visual information available in a CAV.
	2. All sensed spatial information (e.g., GNSS, odometry).
	3. All sensed environmental data (e.g., weather, temperature, air pressure, ice and water on the road, wind, fog etc.)
3. Each Visual Object in the BWR shall be described by
	1. Timestamps.
	2. ID.
	3. State (Static, Dynamic, Unknown).
	4. Position attributes (position, velocity, and acceleration).
	5. Orientation attributes (orientation, velocity, and acceleration).
	6. Semantics (e.g., another CAV or other objects).
	7. The estimated attribute accuracy.
4. The BWR should include a road topology graph describing traffic signalisation, e.g., roads and lane geometry, and lane-specific traffic rules.
5. Each Audio Object in the BWR shall be described by
	1. Its ID.
	2. Its State.
	3. Its physical characteristics, e.g., static, or dynamic.
	4. Its semantics (e.g., other CAVs or other objects).
	5. The estimated accuracies of object parameters.
6. The BWR should allow for easy validation of a trajectory.
7. The BWR shall allow for easy refinement based on new WRs.
8. The BWR shall have a scalable representation, i.e., allowing for:
	1. Transmission of part of the BWR based on required Level of Detail.
	2. Increase of the Environment complexity.
	3. Fast access to critical data.
	4. Fast access to data required by different AIMs.
	5. Incremental refinement of Object and Scene.
	6. Updates for Object and Scene from one snapshot to another.
	7. Deliberative and reactive actions.
9. The BWR shall associate Audio objects with Visual Objects whenever possible.

### FWR requirements

The FWR requirements are:

1. The FRW shall be an extension of the BWR.
2. The FWR shall indicate
	1. Where the ego BWR has major discrepancies with the BWR received from other CAVs.
	2. The identity of the CAV with which the discrepancy has been detenced.
3. The FWR shall include all available information from other CAVs, that enable a CAV to define a Path in the Decision Horizon Time.
4. The objects in the FWR may include Flags (e.g., warning).

**To Respondents**

Respondents are requested to propose a Basic World Representation data format satisfying the requirements. Proposals with justified extended requirements will be considered.

## GNSS Coordinates

**To Respondents**

Respondents are requested to provide a format for the coordinates and the accuracy of the data.

## GNSS Data

Global Navigation Satellite Systems (GNSS) provide spatial information with different accuracies. GNSS can only be relied on when reception conditions are above a certain level. This excludes GNSS in tunnels or urban canyons.

Some data formats are:

1. GPS Exchange Format (GPX) provides an XML schema providing a common GPS data format that can be used to describe waypoints, tracks, and routes.
2. World Geodetic System (WGS) includes the definition of the coordinate system's fundamental and derived constants, the ellipsoidal (normal) Earth Gravitational Model (EGM), a description of the associated World Magnetic Model (WMM), and a current list of local datum transfor­mations.
3. International GNSS Service (IGS) SSR is a format used to disseminate real-time products to support the IGS (igs.org) Real-Time Service. The messages support multi-GNSS and include corrections for orbits, clocks, DCBs, phase-biases and ionospheric delays. Extensions are planned to also cover satellite attitude, phase centre offsets and variations and group delay variations.

**To Respondents**

Respondents are requested to propose a single GNSS data format that is capable to represent the features of all GNSS types.

## Lidar Data

Radio Detection and Ranging (RADAR), LiDAR and Ultrasound are active sensors based on “time-of-flight”, i.e., they measure distance and speed based on the time it takes for a signal to hit an object and be reflected back.

Unlike Radar, however, it operates in the µm range – ultraviolet, visible, or near infrared light. It sends an electromagnetic signal and receives the reflected signal back. These are the features of a typical eye-safe LiDAR:

1. Has a frequency of ~200 THz and a wavelength ~1.5 µm (the visible range is 0.4 to 0.75 µm).
2. Measures the range in each pixel (called also voxels).
3. Pixel grayscale is measured by the intensity variation of the reflected light.
4. The colour of an object can be measured by using more than one wavelength.
5. Velocity can be measured using the Doppler shift in frequency due to motion, or by measuring the position at different times.
6. Micro-motion can be measured using the Doppler shift measured with a coherent LiDAR.
7. Produces 100 kpoints/frame or 1.35 Mbytes: 32\*3 bits (coordinates) +16 bits (ref­lectance). Today 200 kpoints/frame are reasonable.
8. Angular resolution is 0.1º and the vertical field is 40º.
9. A Lidar scan captured at 25 fps generates 270 Mbit/s, i.e., 33.75 Mbytes/s.

The LAS (LASer) format is a binary file format for LiDAR point cloud data specified by the American Society for Photogrammetry and Remote Sensing (ASPRS).

Pcap isa well-established data format for Lidar scans. Other formats, such as E57 is one of them.

**To Respondents**

Respondents are invited to provide a LiDAR data format that facilitates identification, tracking and digital representation of objects having the goal to produce Visual Objects and Scene (Lidar) as required by 6.16.

## Microphone Array Audio

Microphones are used to capture the external sound for the following purposes:

1. Extract speech addressed to CAV.
2. Create Audio Objects and Scene for use in Basic and Full World Representation.
3. Noise suppression inside the passenger cabin.

MPAI-CAE specifies Interleaved Multichannel Audio.

**To Respondents**

Respondents are requested to comment on the usability of the specified technology for MPAI-CAV and/or propose an Audio Array Format suitable to create a 3D sound field representation of the Environment for the stated purposes.

## Moving Object Tracker Data

Moving Object Tracker receives the Visual Objects and Scene data from the different sources – Lidar, Radar, Cameras, Ultrasound, Environment Sound – and provides a list of Visual Objects where each Object has the following associated data:

1. Spatial coordinates
2. Bounding Boxes
3. Coordinated of vertices of Bounding Boxes
4. State
5. Accuracy of the data provided.

**To Respondents**

Respondents are requested to propose a format for the Objects and their list that is friendly to the Basic World Representation format.

## Offline Maps

An Offline Map or HD maps or 3D maps is a roadmap with cm-level accuracy and a high environmental fidelity reporting the positions of pedestrian crossings, traffic lights/signs, barriers etc. at the time the Offline Map has been created.

Worth noting are:

1. Navigation Data Standards calls itself “The worldwide standard for map data in automotive eco-systems”. Their NDS specification covers data model, storage format, interfaces, and protocols.
2. SharedStreets Referencing System is a global non-proprietary system for describing streets.

**To Respondents**

Respondents are requested to propose an Offline Map Format. The Format should support different levels of conformance.

## Radar Data

Radar operates in the mm range. It can detect vehicles (CAVs and trucks) because they typically reflect Radar signals while objects that are smaller and have less reflectance, e.g., pedestrians and motorcycles have a poor reflectance. In a busy environment, the reflections of big vehicles can overcome a motorcycle’s causing missed detection of important objects (e.g., a child next to a vehicle), while a can may produce an image out of proportion to its size.

The main features of Radar are:

1. Measures distance.
2. Is independent of environment.
3. Provides low resolution images.
4. Provides distance (short range radar in the 25 GHz band).
5. Detects objects and measures speed @ ≤ 250 m (long range radar in the 76-77 GHz). Typical long-range radar systems have ranges of 80-200 m. Small antenna because wavelength is ~3.5-4 mm. Atmospheric absorption limits interference with other systems. A multitask 94-GHz pulse Doppler radar has 25-cm radial and 1.5 degrees angular resolution

Radar sensors build a representation of the environment based on the observation of complex, scattered radio waves, from which information of an object’s distance and velocity can be derived.

Known Radar data formats include:

1. OPERA BUFR format.
2. hdf5 formats.
3. NetCDF files generated by the commercial EDGE software.
4. hdf5 files generated by the commercial GAMIC software.
5. German Weather Services quantitative local scan format (DX).
6. Quantitative composite format (RADOLAN, see German Weather Service, 2004).

**To Respondents**

Respondents are invited to propose a format of Radar images that facilitates identification, tracking and representation of objects having the goal to produce Visual Objects and Scene (Radar) as required by 6.17.

## State

State is the set of the following CAV attributes at a given time:

1. Pose, Velocity and Acceleration
2. Orientation, Angular Velocity and Angular Acceleration.

**To Respondents**

Respondents are requested to propose a State Format suitable for use in CAVs.

## Traffic Signalisation

Traffic Signalisation types are:

1. Traffic signs.
2. Road signs.
3. Placement signs.
4. Acoustic signs.
5. Traffic lights.

They are recognised from camera, lidar, rad, ultrasound images. Possible classification as below (From: https://www.researchgate.net/figure/Traffic-Signals-ontology\_fig4\_271376130).



**To Respondents**

Respondents are requested to propose a set of Traffic Signalisation Descriptors.

## Ultrasound Data

These are the main features of Ultrasound:

1. To operate at frequencies above 20 kHz.
2. To be independent of environment.
3. To yield images with low resolution.
4. To work on a limited range (≤ 10 m)

The Ultrasound File Format initiative has defined the Ultrasound File Format (UFF) format [30].

**To Respondents**

Respondents are invited to propose an Ultrasound Format that facilitates identification, tracking and representation of sound objects with the goal to produce Visual Objects and Scene (Radar) as required by 6.13.

## Ultrasound World Representation (Ultrasound)

The expected output is a scene representation used by Moving Object Tracker, Traffic Signalisation Recogniser and Basic World Representation.

**To Respondents**

Respondents are invited to provide a Format for scenes captured by Ultrasound. The format should be sufficiently generic to be capable to be used – or adapted for use – in scenes captured by Lidar, Radar and Video devices.

## Video Camera data

The output of the video camera is used to produce Visual Objects and Scene (Camera).

**To Respondents**

Respondents are invited to provide a data Format for RGB-D cameras.

## Visual World Representation (Camera)

The expected output is a scene representation used by Moving Object Tracker, Traffic Signalisation Recogniser and Basic World Representation.

**To Respondents**

Respondents are invited to provide a Format for scenes captured by cameras. The format should be sufficiently generic to be capable to be used – or adapted for use– in scenes captured by Radar, Lidar and Ultrasound devices.

## Visual World Representation (Lidar)

The expected output is a scene representation used by Moving Object Tracker, Traffic Signalisation Recogniser and Basic World Representation.

**To Respondents**

Respondents are invited to provide a Format for scenes captured by Lidars. The format should be sufficiently generic to be capable to be used – or adapted for use– in scenes captured by Radar, Video and Ultrasound devices.

## Visual World Representation (Radar)

The expected output is a scene representation used by Moving Object Tracker, Traffic Signalisation Recogniser and Basic World Representation.

**To Respondents**

Respondents are invited to provide a Format for scenes captured by Radars. The format should be sufficiently generic to be capable to be used– or adapted for use – in scenes captured by Lidar, Video and Ultrasound devices.

# Terminology

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Accuracy | The precision with which a Data Type measures an entity. |
| Audio object | An aurally identifiable sound source. |
| Data Format | The standard digital representation of Data. |
| Data Type | An elementary component of a Data Format. |
| Environment Sensing Technology | One of the technologies used by the Environment Sensing Subsystem to sense the Environment. |
| Format compatibility | The ability of a Data Format to be losslessly converted to another Data Format. |
| Visual object | A visually identifiable set of pixels/voxels. |
| World Representation | A digital representation of the Environment produced by a CAV Environment Sensing Technology or an Offline Map. |