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

|  |  |
| --- | --- |
|  | |
| **N380** | 2021/09/30 |
| **Source** | Requirements (OSD) |
| **Title** | MPAI-OSD report and plans |
| **Target** | MPAI Members |

# Introduction

Visual object and scene description (MPAI-OSD) is an MPAI project at the Use Case stage, collecting Use Cases sharing the goal of describing visual objects and, in some cases, locate them in the space.

By scene description we mean the usual description of objects and their attributes in a scene and the semantic description of the objects.

AIMs in the MPAI-OSD area have already been requested in Conversation with emotion and Mul­timodal Conversation. However, no specific responses have been received.

New use cases are constantly identified that require new AIMs falling under the MPAI-OSD scope.

# Description of Use Cases

## Audio Tape Irregularity Detection (MPAI-CAE)

This belongs to the family of generic object description.

MPAI is using this component in the MPAI-CAE Use Case Audio Recording Preservation.

It is designed to:

1. Receive the video signal of a camera pointing to the magnetic reading head of a traditional audio tape.
2. Detect the images that show irregularities on the tape.
3. Provide as output, if an image shows an irregularity:
   1. The image
   2. The type of irregularity
   3. The time code

## Identify object in a human’s hand (MQA)

This belongs to the family of generic object description.

MPAI is using this component in the MPAI-MMC Use Case Multimodal Question Answering.

It is designed to:

1. Receive the picture of an object.
2. Recognise the type of object.
3. Provide the object identifier as outputs.

## Detecting emotion and meaning in human face (CWE)

This belongs to the family of human description.

MPAI is using this component in MPAI-MMC Conversation with Emotion Use Case which is designed to:

1. Receive a video of the face of a human.
2. Identify the type and intensity of the emotion in the face of the human.
3. Provide as output:
   1. The type of emotion out of a finite set of codified emotions.
   2. The intensity (grade) of the emotion.
   3. The time stamp that the type and intensity of emotion refers to.

## Visual and scene for Connected Autonomous Vehicles (CAV)

This contains many elements belonging to human, animal, vehicle, road sign and traffic light description. There is a large variety of sensing devices:

1. 2D and 3D cameras
2. Lidar
3. Radar
4. Ultrasound

and other sources of information such as odometer and GSSN to create a Basic World Represen­tation (BWR). A CAV exchanges its BWRs with other CAVs in range and produces a refined Full World Representation (FWR).

MPAI-CAV Human-CAV Interaction Use Case where CAV needs to

1. Detect the emotion of a passenger to be able to have better conversations or provide better responses to queries.
2. Locate passengers in the compartment so that the avatar representing the CAV can gaze at them in a more natural way.

## Avatar-based videoconference (MCS)

Geographically distributed users can send their data to a virtual space and create local 3D audio-visual spaces where they can see a virtual meeting populated with avatars whose face and head are animated that they can navigate vithout moving their avatar.

1. MPAI-MCS Local Avatar Videconference Use Case where a participant in a videoconference is represented by an avatar whose torso is faithfully represented by an avatar.
2. MPAI-MCS Virtual eLearning where teacher and students are represented by avatars and can interact with 3D Audio-Visual Objects Enter, navigate and act on 3D audio-visual objects by doing:
   1. Define a portion of the object – manual or automatic
   2. Count objects per unit volume
   3. Detect structures in a (portion of) the 3D AV object
   4. Combine objects
   5. Call an anomaly detector on a portion with an anomaly criterion
   6. Follow a link to another portion of the object
   7. 3D print (portions of) the 3D AV object

## Tracking video game player’s movements

This Use Case belongs to the family of human description.

It is a system designed automatically to understand the game player’s physical movements in a video game. The features of the movements are:

1. The human object is largely static, and only hand/arm and finger movements are detected.
2. The types of movements are limited in number.
3. The system should understand the movements fast and accurately.

The system is designed to:

1. Receive a video.
2. Compute descriptors of the human.
3. Understand the inten­tion expressed by the movements from the descriptors.

## Correct Posture

This Use Case belongs to family of human description.

It is a system designed to advise the user by suggesting how they should correct their pose.

The main features of this Use Case are

1. The human using the application walks in a restricted environment.
2. Very specific type movements must be detected with high accuracy.
3. The detected movements are compared with reference movements.

The system is designed to:

1. Receive a video.
2. Compute descriptors of the human.
3. Compares the descriptors with reference descriptors.
4. Provide suggestions about movement corrections.

## Integrative genomic/video experiments (animals)

This belongs to the family of animal description (however, see later).

MPAI is using this component in several MPAI-GSA Use Cases.

It is a system designed to

1. Receive a sequence of images containing laboratory animals with specified genomic data of which the effects on behavioural patterns are assessed.
2. Compute behavioural patterns of living organisms, e.g., measures of the parameters of animal activity, everywhere and/or in specified ROI, such as:
   1. Distance.
   2. (average) velocity.
   3. Acceleration.
   4. Time spent.
   5. Time spent near walls.
   6. Trajectories.
   7. Turning speed.

There is another component that belongs to the family of plant description.