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

|  |  |
| --- | --- |
| **Public Document** | |
| N93 | 2020/12/16 |
| Source | General Assembly #3 (MPAI-3) |
| Title | MPAI Application Note #8 - MPAI-OSD – Visual Object and Scene Description |
| Target | MPAI Members |

**Proponents**: Leonardo Chiariglione (CEDEO), Luigi Troiano (Kebula)

**Description**: This use case addresses the “object and scene description” component of several use cases considered in MPAI’s use case document rev2.0 (N46). “Object and scene description” is used to indicate a descrip­tion (language) of objects and their attributes, and the semantic description of the individual objects in a scene.

**Comments**:

The Examples section of this use case shows that proprietary solutions can address the needs of the examples. However, proprietary solutions have the following disadvantages

1. Lack of interoperability, i.e. proprietary solutions create silos
2. Adoption of a single technology as a black box
3. Closed applications without access to the enabling technology
4. Delegation of enabling technology innovation to the technology provider (lock-in).

On the other hand, a standard representation of the objects in a scene and of the scene allows for

1. Interoperable applications, i.e. data can be moved from one domain to another
2. Possibility to select the enabling technology with the only constraint of preserving the interfaces
3. Possibility to create open applications where the enabling technology can be replaced safeguarding interoperability of data
4. MPAI, by decision of its members, can decide to extend an existing standard or to create a new one when progress of technology – independently achieved – requires it.

**Examples**:

1. *Multiplayer online gaming*

In the Hide and seek example of the “Distributed multiplayer online gaming in Next Generation Games” use case (ME.MP-09)

1. Player A
   1. Points their smartphone to a scene populated by persons and other objects
   2. Sends the description of the scene and their objects to a server
   3. Sends commands to the server to hide a synthetic person among the real persons
2. The server
   1. Understands the scene description received
   2. Adds the animated synthetic person to the scene
   3. Distributes the composite scene to all players
3. All players but player A seek the synthetic person in the crowd.

The enabling technologies are

1. Description of the persons in the scene and their movements
2. Description of other objects
3. Description of the position of all objects in the scene
4. Transmission of all descriptors to the server
5. Transmission of commands to animate a synthetic person (player trying to hide)
6. Animation of a new object (person trying to hide)
7. Creation of a new scene description containing
   1. The objects received
   2. The new animated object
   3. Any other object that the logic of the game may require
8. Transmission of new scene description to all players

In summary:

1. Visual objects and scene description
2. Animation of objects described by 1.
3. *AI-assisted driving*

In the “AI-assisted driving” (TP.MP-01) use case, a car computes the descriptors of the scene and displays appropriate messages to the driver. When considered appropriate, the car communicates the scene descriptors to the neighbouring cars.

At first glance, the technologies needed in usage example #1 are the same that are needed in this usage example, as it can be seen from the workflow below

1. Camera of the car captures the scene of a human crossing the road
2. Car (or camera) computes the visual descriptors, the position in space, speed and direction of movement
3. Car assesses the urgency and display appropriate information of the human to driver and, if needed, to nearby vehicles
4. *Vision-to-sound transformation*

In the Vision-to-Sound Transformation (HC.AV-01) use case, a 3D Visual space is transformed into a 3D Audio space. An approach to solve this problem is based on a full description of the visual scene and its objects and then on a standard visual-to-audio conversion.

The requirements of this usage example are not known. Is the scene or its objects static or dynamic?

In either case, however, the potential Visual Object and Scene Description standard could be used.

1. *Tracking video game player’s movements*

In the “Tracking game player’s movements” (ME.MP-12) use case, the client sends descriptions of the game player’s movements to a server. The server decodes the game player’s move intention from the descriptors.

The technology in this usage example is akin to the preceding ones. However, in this case

1. The human object is largely static, only hand/arm, fingers movements are detected
2. The types of human object movements are more restricted
3. High accuracy is required
4. *Correct Pose*

In the “Correct Posture” (HC.MP-02) use case, an AI application analyses the video of a person to suggest how the person should correct their posture.

The technology in this usage example is similar in terms of movement representation accuracy as the preceding one. However, in this case

1. The human object walks in a restricted environment
2. Very specific type movements must be detected with high accuracy
3. Scene description is typically not required.
4. *Person matching*

In the “AI-services for next generation TV” (ME.MP-11) use case, descriptors of the video prog­ram and associated metadata reach the set top box together with the actual program.

When the user wants to know more about an object:

1. User points the remote control to the object on the screen
2. Set top box
   1. Computes object descriptors
   2. Compares the computed descriptors with those in the stream
   3. When a match is found, executes as per metadata

This usage example differs from the preceding ones because the scene and its objects are 2D, while in the other usage examples the scene and the objects can be considered as 3D.

There has been and is significant research on 2D scene descriptors for searching objects in a data base of TV programs.

1. *Integrative genomic/video experiments*

In the “Integrative analysis of multi-source genomic/sensor experiments” (ST.OD-06) use case, an AI-based application assesses genomic data and their effects on living organisms. An example is the study of movements of Zebrafish, a (sub-)tropical fish that is widely used in this kind of tests. Currently used programs provide: speed, average speed, acceleration, time spent in ROI, trajectories, identification of same animal in different videos, turning speed, time near walls and more [1].

1. *Audio Recording Preservation*

In this Use Case, features of images taken from the video of an audio tape passing in front of the magnetic head are used to search in a Knowledge Base of magnetic tape irregularities.

1. *Conversation with emotion*

In this use case a human has a conversation with a machine. The machine takes pictures of the human, extracts features and queries a Knowledge Base of features of known emotions

1. *Multimodal Question Answering*

In this use case a human interrogates a machine using a picture of an object. The machine extracts features of the image and queries a Knowledge Base of features of objects.

This usage example can be satisfied by a standard that simply defines the syntax and semantics of the output of AIMs such as speed, average speed, acceleration, time spent in ROI etc. or it could dig into the low-level descriptors that are used to compute the said high-level descriptors.

**Requirements:**

This is an initial set of requirements

1. The standard shall enable the description of human objects to enable
   1. Animation of human models
   2. Interpretation of application-specific gestures
2. The standard shall enable the description of scenes where
   1. Individual objects can be manipulated
   2. The scene can be updated

**Object of standard**:

1. Low-level descriptors of selected objects
2. Application-dependent high-level descriptors
3. Description of dynamic scenes

Based on the examples above, MPAI-OSD could define target several standards covering specific applications

|  |  |
| --- | --- |
| **Potential standard** | **Example** |
| Person movement description | Multiplayer online gaming  AI-assisted driving  Correct pose  Tracking game player’s movements |
| Scene description | Multiplayer online gaming  AI-assisted driving  Integrative genomic/video experiments |
| Generic object description | Multiplayer online gaming  AI-assisted driving |
| Person identification | Person matching |
| Animal movements description | Integrative genomic/video experiments |

**Benefits:** The standard would stimulate a competition among developers of neural networks that can best convert objects into the required standard set of des­crip­tors.

**Bottlenecks**: many applications would benefit from the availability of independently sourced object description solutions that expose the MPAI interfaces. Decoupling of the enabling technologies from the applications would benefit both sides

**Social aspects**: the impact of the standard is already being assessed by the many examples of highly realistic synthetic faces

**Success criteria**: the number and success of applications that use the standard

**References:**

[1]