MPAI - Artificial Intelligence for Industrial Standards

Giorgio Audrito[1], Andrea Basso[2], Davide Cavagnino[1], Leonardo Chiariglione[3], Ferruccio Damiani[1], Attilio Fiandrotti[1], Roberto Iacoviello[4], Maurizio Lucenteforte[1], Marco Mazzaglia[2], Alessandra Mosca[5], and Gianluca Torta[1]

[1]Dipartimento di Informatica, University of Turin, Turin, Italy
[2]Synesthesia s.r.l., Turin, Italy
[3]MPAI – Moving Picture, Audio and Data Coding by Artificial Intelligence, Geneva, Switzerland
[4]RAI – Radiotelevisione italiana, Rome, Italy
[5]Sisvel Technology, None Torinese, Turin, Italy

Abstract

This document describes the contribution of the University of Turin Computer Science Department to the standardization activities of MPAI (Moving Picture, audio and data coding by Artificial Intelligence), an international organization whose goal is the development of industry standards using technologies based on Artificial Intelligence.

1 Introduction to MPAI

The University of Turin (UniTo) is a member of MPAI, an international, unaffiliated, non-profit standards developing organisation has the mission to develop Artificial Intelligence (AI) enabled data coding standards. MPAI has defined and adopted a standard development process that improves upon other standardization bodies shortcomings, including lack of a clear Intellectual Property Rights (IPR) licensing framework. The main targets of MPAI standardisation are AI modules (AIMs), processing elements whose function, input and output data syntax and semantics, but not the internals, are normatively defined; and structured aggregations of AIMs called AI Workflows (AIWs). They are executed in an MPAI defined environment called AI Framework (AIF). MPAI-specified Use Cases are executed as AIWs in the AIF. By basing its standards on the AIM-AIW-AIF model, MPAI can reuse standard components. This is already true for two of the developed standards, e.g., emotion and speech features, and is apparent for many of those under development. More importantly, AIM implementers can have a low entry barrier to an open competitive market for their implementations: Applications (i.e., AIW) implementers can find the AIMs they need on the open market and make their applications available to the Store. Consumers have a wider choice of better AI applications from competing application developers and can select the interoperability level that is convenient to them, tested by the Ecosystem players. Innovation is fueled by the demand for novel and better performing AIMs. In just a year, MPAI has published 5 standards and seven more standards are under development. [1] MPAI is the root of trust of an ecosystem offering users access to the benefits of reliable, robust, fair and replicable AI and a guarantee of increased transparency, trust and reliability as the MPAI-defined Interoperability Level of an Implementation moves from 1 to 3. In the meantime, MPAI will extend its standards and produce novel ones serving new industries and users. UniTo computer science is actively contributing to four different projects, as detailed below

The AIF group defines the infrastructure that enables the execution, control and management of AIWs and AIMs. The MPAI-AIF Technical Specification specifies architecture, interfaces, protocols and Application Programming Interfaces (API) of the AI Framework, especially designed for execution of AI-based AIMs, but also suitable for mixed AI and traditional data processing workflows. It is therefore a foundational standard, on which all the other MPAI application standards are built. Currently, version 1 of the MPAI-AIF standard has been approved.

CAV is a Requirements subgroup developing the document called Use Cases and Functional Requirements. The document describes 1) the 4 subsystems of an MPAI-CAV (Human-CAV Interaction, Environment Sensing Subsystem, Autonomous Motion Subsystem and Motion Actuation Subsystem, and the infrastructure that allows a subsystem or its modules to communicate with their peers in other CAVs, and 2) the functional requirements of all data that are expected to be subject to MPAI standardisation.

The EIDOS group [2] takes part in the Server-based Predictive Multiplayer Gaming (MPAI-SPG) and AI-Enhanced Video Coding (MPAI-EVC) projects.

The MPAI-SPG project has the objective to develop a standard to define an online gaming architecture that helps to reduce the audio, video and gameplay discontinuities caused by network problems or cheating clients. This is a very important topic for video-game players who need high performance GPUs and displays for an improved gaming experience [Tamasi, 2019].

MPAI-EVC aims at enhancing the performance of a traditional video codecs (MPEG-5 EVC in this case) by improving or replacing traditional encoding tools with AI-based tools. The MPAI-EVC evidence project seeks to demonstrate that extending or replacing such tools with AI-enabled counterparts can offer at least 25% performance improvement. The

[1] https://mpai.community/standards/workplan/

[2] http://eidos.di.unito.it/

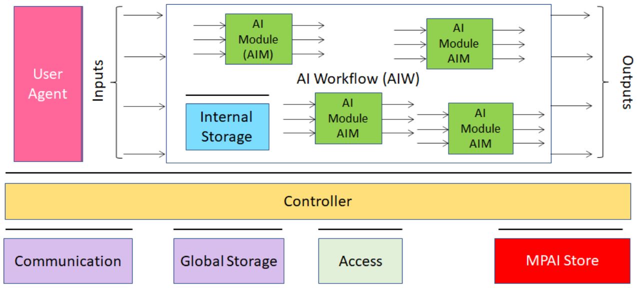

Figura 1: The MPAI-AIF Reference Model.

group contributes to the project its experience in the development of video coding standards and its skills into deep learning based image processing together with its industrial partners Sisvel Technology and RAI.

2 MPAI AIF

The purpose of the AI Framework (AIF) is to support the execution of workflows (AIWs) built from functional components called *AI Modules (AIM)* (AIMs).

The Reference Model of AIF, shown in Figure 1, envisions an entity named MPAI Store handling: submission, registration,identification, validation and access to AIMs and AIWs As shown in Figure 1, AIWs connect AIMs into computational graphs. In the simpler case these are directional and acyclic (DAGs) and express computations that flow from the inputs to the outputs of the workflow. In more complex cases they are cyclic, allowing feedbacks between AIMs.

An important characteristic of AIF, is the adoption of a *zero-trust model*, which advocates mutual authentication of the components of a system, including checking the identity and integrity of components irrespective of location, and providing access based on the confidence on components identity and health.

The overall control of the workflow execution is the responsibility of the *Controller* component of AIF (Figure 1). The standard allows implementations with both centralized and distributed control architectures: the former will typically be preferred when the AIMs are in a limited number, are tightly coupled, and (mostly) local, and/or special security constraints apply. On the other hand, a distributed architecture is generally more effective and efficient when workflows encompass many physically distributed AIMs that are mostly connected with geographically close AIMs.

Independently of the control architecture, the AIF supports some operations that act globally on the workflow, such as *start-up suspend resume shutdown*. Moreover, the *Storage and Access* components of the AIF can be distributed as well.

In some important scenarios, however, different instances of the same workflow (or even of different workflows) must communicate with each other in order to perform their task. A notable example is the CAV-to-Everything interaction (see next section).

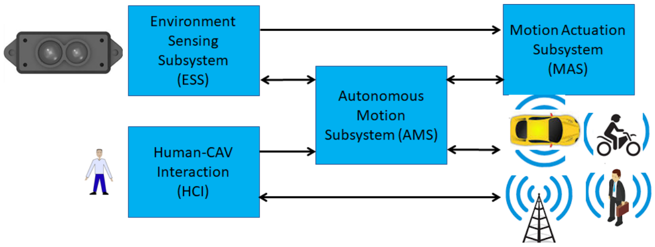
Environment Sensing Subsystem (ESS)
Motion Actuation Subsystem (MAS)
Autonomous Motion Subsystem (AMS)
Human-CAV Interaction (HCI)

Figura 2: The MPAI-CAV Reference Model with its Subsystems.

3 MPAI CAV

Connected Autonomous Vehicles (CAVs) are a lively area for research and experimentation [Elliott *et al.*, 2019]. Standardization of CAV components will be required because of 1) the different nature of the interacting technologies making up a CAV, 2) the sheer size of the future CAV market, and 3) the need for users and regulators alike to be assured of CAV safety, reliability and explainability [Imai, 2019].

Some affected industries may find standardization premature at this time. However, according to the MPAI philosophy, there are reasons to the contrary. Indeed, CAVs belong at best to an industry still being formed, that will target the production of affordable units in the hundreds of millions, the components of which will be produced by disparate sources. Since CAVs promise to have a major positive impact on the environment and society, they are urgently needed and, to reduce their costs, early availability of critical components is required. The role of standards is exactly that of facilitating the achievement of this goal.

The development of standards can be accelerated by creating a CAV Reference Model (RM) proposing components and their interfaces. Progression from research to standardization can then be implemented as a series of transitions between research (proposing components and interfaces to standardization) and standardization (either requesting more results, or refining the results, or adopting the proposal).

MPAI has produced a RM where a CAV is subdivided in four subsystems (Figure 2):

- Human-CAV interaction (HCI), handling interactions with humans;
- Environment Sensing Subsystem (ESS), acquiring information from the physical environment via a variety of sensors;
- Autonomous Motion Subsystem (AMS), generating commands to drive the CAV to the intended destination based on a Full World Representation;
- and Motion Actuation Subsystem (MAS), providing environment information¸ and receiving/actuating motion commands in the physical world.

A fundamental part of the RM concerns the definition of the CAV to Everything (V2X) communications between the CAV and the external world (botom-right objects in Figure 2). In particular, a CAV exchanges information via radio with other entities, including other CAVs in range, Roadside Units, and Traffic Lights, thereby improving its environment perception capabilities. It is worth noting that, to support such communications, the AI Framework described in the previous section defines a specific set of APIs, intended for communication between AIF Controllers hosted on separate physical devices.

4 MPAI SPG

In Massive Multiplayer Online Gaming (MMOG), a very large number of users (clients) play the same instance of a game which is controlled by a centralized server [Glazer e Madhav, 2015]. In general, server and clients are geographically distributed and use heterogeneous networks and protocols to exchange game data (i.e. input commands, game state updates, or even rendered images). At least two approaches for the management of such a system are possible:

- Traditional online gaming: the server receives input commands from clients that allow it to update the game state which will in turn be transmitted to every client; as a consequence a client, starting from the received game state, will be able to render the related video frame.
- Online cloud gaming: the server updates the game state as in the previous approach, but here the video frames are rendered and transmitted to the clients using cloud servers (i.e. Google Stadia [Google, 2022]).

The problem of cheating in video games consists of one or more clients using mechanisms to gain an advantage over the fair execution of a game match. A possible answer to this problem is the adoption of an authoritative server, which is the only entity maintaining and updating the game state, while no client can update it, preventing the aforementioned cheating attempts.

In online gaming it is also recommended to ensure an adequate gaming experience even in network critical conditions, where transmission errors cause packet delays or loss of data.

In order to mitigate the previously cited problems MPAI-SPG proposes a game architecture including a digital twin of the game server that may be implemented using Artificial Intelligence techniques and trained with Machine Learning approaches: the task of this entity is to compensate for any client data loss or delay and to detect possible client cheating attempts by predicting the current game state based on previous ones.

Our implementation of the game server digital twin (hereinafter referred to as SPG) is realized with a number of interconnected deep neural networks able to predict the next game state from a sequence of previous game states.

The learning phase is fed using the data (game state sequences) extracted from a reasonable number of executions of game matches, which can be played by humans or by AI agents acting as game players. The latter solution allows the automatic generation of an adequate number of consistent game state sequences to be used for learning. Such agents are trained with reinforcement or imitation learning approaches.

The function of SPG is to infer the possible subsequent states of the game, which can be used by the authoritative server when client data is missing or has an unexpected value. In the former case SPG is used to integrate the missing data, while in the latter cheating attempts can be detected and dealt with.

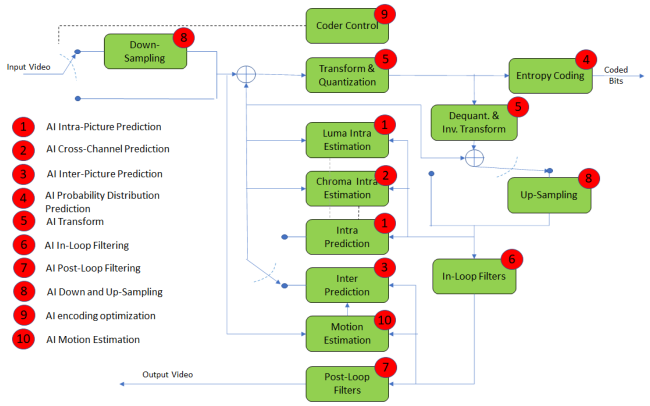

Figura 3: The MPAI-EVC model and main encoding tools.

Given the sequential nature of the data to be handled, the most suitable architectures are from the class of Recurrent Neural Networks (RNN), in particular Long Short-Term Memory (LSTM) which are characterized by feedback connections to take into account the time evolution of the game.

5 MPAI EVC

Since the day MPAI was announced, there has been considerable interest in the application of AI to video coding and that resulted in the creation of the EVC project. Video contents nowadays accounts for more than 70% of Internet traffic volume[CISCO, 2021], hence the interest into efficient video coding technologies able to cope with tomorrow bandwidth-demanding video services (4K video, immersive contents, etc.). Existing video coding standards used in Internet streaming or aerial broadcasting over the air or cable rely on a clever combination of hand-designed encoding tools, each bringing its own contribution to the overall codec performance. The EVC project aims at enhancing the performance of existing video coding tools by replacing selected encoding tools with AI-based counterparts. The MPEG EVC video coding standard has been selected as a starting point because its Baseline Profile is made up with 20+ years old technologies and has a compression performance close to HEVC, while the performance of the Main Profile exceeds that of HEVC by about 36%. Additionally, some patent holders have announced that they would publish their licence within 2 years after approval of the EVC standard. Fig. 3 shows the internal architecture of the EVC codec highlighting the main coding tools it relies upon. Up to this point, two tools have been investigated by the EVC project, namely the intra prediction tool and the upsamplig tool.

The first tool we investigated is the intra prediction tool. In modern video codecs, intra-frame prediction removes the spatial correlation within the same picture generating a predictor for the image block to be encoded by extrapolating the pixel values from a previously decoded neighbourhood. The predicted block is the subtracted from the original block to be included, producing a residual block that is transformed, quantized and entropy coded as other image blocks before being inserted into the bitstream. The rationale behind intra prediction is that encoding the difference between the block to be coded and a predictor requires fewer bits than encoding the original block. Namely, the better the predctor, i.e. the closer to the block to be encoded, the lower the energy of the encoded residual and the higher the coding efficiency. In EVC, intra prediction consists in proposing a set of 5 predefined functions and choosing the best among them in a rate-distortion sense, posing a limitation of the possible number of functions. However, AI has the potential to approximate complex functions in predictive tasks such as the generation of a future block given an input's context. In this sense, we addressed the problem of predicting a block given its context as an image inpainting problem, i.e. recovering pixels of an image that are unavailable due to, e.g. occlusions or information loss. Recently, deep convolutional generative neural networks have shown to outperform existing image inpainting methods thanks to their ability to learn complex non linear functions. Namely, masked convolutional neural networks have been recently proposed for image inpainting exploiting the apriori information on missing pixels to weight out such pixels from the context used to recover the missing image area. The method we propose relies on masked convolutions to generate the block predictor starting from a decoded context of 64×64 pixel. In detail, we replaced EVC predictor mode 0 (i.e., the DC mode predictor) with a novel predictor that is computer by a masked convolutional autoencoder for each block to be encoded. The masksed autoencoder is trained in a weakly supervised way extracting random patches from about 800 images representing various types of contents. Our encoding experiments with a of HD images commonly used for similar tests has shown preliminary gains in the 4 5% range. Future challenges include modeling the encoding rate and controlling the complexity of the convolutional autoencoder so that it can be implemented in FPGA.

The second tool we investigated is the super resolution tool. Super Resolution creates a single image with two or more times the linear resolution e.g., the enhanced image will have twice the width and twice the height of the original image, or four times the total pixel count. For the Super Resolution track a state-of-the-art neural network (Densely Residual Laplacian Super Resolution) was selected because it introduces a new type of architecture based on cascading over residual, which can assist in training deep networks. A dataset to train the super resolution network has been created with 3 resolutions (4k, HD, and SD), 4 values of picture quality, two coding tool sets (deblocking enabled, deblocking disabled) for a total of 170 GB dataset. The super resolution step was added as a post processing tool i.e., the picture before encoding with EVC baseline profile was downscaled and then the super resolution network was applied to the decoded picture to get the native resolution. Many experiments have been performed to find the right procedure to select a region in the picture (crop), i.e., an objective metric to choose one or more crops inside the input picture in such a way that a trade-off between GPU memory and compression performance is achieved. Our encoding experiments on the SD to HD task has shown preliminary gains of about 5%.

Additional tools considered for further experiments by the MPAI-EVC projects include:

- in-loop filtering: reduce the blockiness effect by filtering out some high frequencies caused by coded blocks.
- motion compensation: use Deep Learning architectures to improve the motion compensation.
- inter prediction: estimate the motion using Deep Learning architectures to refine the quality of inter-predicted blocks; introduce new inter prediction mode to predict a frame avoiding the use of side information.
- quantization: use a neural network-based quantization strategy to improve the uniform scalar quantization used in classical video coding because it does match the characteristics of the human visual system.
- arithmetic encoder: use neural networks to better predict the probability distribution of coding modes.

6 Conclusions

UniTo is deeply involved in MPAI, an environment where projects in different areas sharing the goal of using Artificial Intelligence to encode data are submitted. The areas presented in this paper cover just a part of the activities: Context-based Audio Enhance-ment, Multimodal Communication, Company Performance Prediction are examples of other standardisation areas. Future work will include 1) an implementation of the AI Framework (AIF), 2) integration of AIF with the Connected Auronomous Vehicle (CAV) to test different strategies to optimise the movement of swarms of CAVs or drones, 3) experimenting with different neural network architectures to improve the performance of SPG, 4) performance enhancement of MPEG-5 EVC with AI-based tools with the goal to reach a total 25% improvement, test the achieved results and move to the next phase of standard development.

Riferimenti bibliografici

[CISCO, 2021] CISCO. Global 2021 forecast highlights. https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf, 2021.

[Elliott *et al.*, 2019] David Elliott, Walter Keen, e Lei Miao. Recent advances in connected and automated vehicles. *Journal of Traffic and Transportation Engineering (English Edition)*, 6(2):109–131, 2019.

[Glazer e Madhav, 2015] Josh Glazer e Sanjay Madhav. *Multiplayer game programming: Architecting networked games*. Addison-Wesley Professional, 2015.

[Google, 2022] Google. Stadia. https://stadia.dev/intl/en_uk/about, 2022.

[Imai, 2019] Takeyoshi Imai. Legal regulation of autonomous driving technology: Current conditions and issues in japan. *IATSS Research*, 43(4):263–267, 2019.

[Tamasi, 2019] Tony Tamasi. Why Does High FPS Matter For Esports? https://www.nvidia.com/en-us/geforce/news/what-is-fps-and-how-it-helps-you-win-games/, 2019.