VERSION 2.0

OCT 17TH, 2022.

MULTIMODAL CORRESPONDENCE BETWEEN SELF-SUPERVISED REPRESENTATIONS

IPTC-AMAZON COLLABORATION

INFORMATION PROCESSING AND TELECOMMUNICATIONS CENTER
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN
UNIVERSIDAD POLITÉCNICA DE MADRID

MULTIMODAL CORRESPONDENCE BETWEEN SELF-SUPERVISED REPRESENTATIONS

This document presents the main research activities.

MAIN RESEARCH AREA

The IPTC-Amazon collaboration will have a duration of 1 year and it will be focused on developing technologies to extract and combine self-supervised representations for multimedia processing. These technologies have a big potential in many areas such as content generation (audio, image, video, or sign language representation), classification, labelling or searching.

ACTIVITY PLAN

The activity plan proposed for the first year is composed of these activities:

1.- Collaboration presentation and diffusion

The IPTC-Amazon collaboration will organize several events/meetings to present this activity plan to all the students at the ETSI Telecomunicación (UPM). This program will be presented during regular classes in higher degree courses and in the first year of master's degrees. The target is to recruit the best students for this plan

2.- Students formation

After student recruitment, professors involved in this collaboration will provide a team of 5 selected students with the necessary technological skills to the develop several applications. The main aspects to be considered are:

- Introduction to deep learning strategies like CNN, RNN, LSTM, Transformers, Adversarial and Contrastive learning, etc.
- Use and installation of different developed tools for deep learning. Special attention will be made on contrastive learning tools like CLIP or LIP and variants like WavCLIP, AudioCLIP or MotionCLIP. Also, video and audio processing tools (like OpenPose, MediaPipe, etc.) will be considered.
- As a result of this training process, the team of students will develop a framework where they can experiment adding embedding audio or image representations from pre-trained models into existing recent proposals of shared embedding spaces for images and text such as CLIP or LIT.
- Using this framework with different datasets, students will explore and evaluate several downstream applications such as multimodal classification, tagging, retrieval and generation.

3.- Applications development

After the formation step, students will address the development of several applications. Some examples of possible applications are:

- Sign language motion generation from high level sign characteristics
- Speaker diarization with multimodal inputs
- Pose and spatial movement as input for dynamic content search & generation
- Entangling AI-audio synthesis models and multimodal representations
- Zero-shot sonorizing of video sequences

These applications could change depending on the state-of-the-art evolution and the availability of the necessary resources.

4.- Students supervision

The students' supervision will be carried out considering these activities:

- Every student will be assigned a supervisor, a PhD professor with experience with these technologies. This professor will hold supervision meetings every week.
- Additionally, joint meetings and sessions will be organized monthly where the students will present their last achievements. Several professors will attend these meetings to provide feedback and give advice about the next steps.

5.- Results

The expected results will be:

- All the students will write a detailed report describing all the analyses and experiment carried out during the year.
- At least, 3 prototypes and demonstrations will be developed to show the main research achievements.
- 3 papers will be submitted to international conferences or journals.

APPLICATIONS DEVELOPMENT

This section describes several examples of possible applications. The final developed applications could vary depending on the state-of-the-art evolution and the availability of the required resources.

SIGN LANGUAGE MOTION GENERATION FROM HIGH LEVEL SIGN CHARACTERISTICS

Using avatars for sign language has an important flexibility because you do need to record a person signing every specific content. On the other side, controlling an avatar can be difficult because it is necessary to control all its elements (skeleton and mesh). This difficulty can produce movements with low naturality. The proposed application focuses on generating natural sign language motion information (in 2D) from high level sign characteristics (hand shape, orientation, hand localization, etc.). The main technical target would be to develop a deep learning algorithm able to associate high level sign characteristics to skeleton motion. One important limitation is the availability of enough data to develop and train a good DL system. This research target is planned in two main phases:

Phase 1: Generating a good dataset containing a relevant number of signs descriptions with motion information from several performances of the same sign. In this phase, we plan the following activities:

1. Obtain a big parallel corpus including sign characteristics and video sequences. Currently, we have a small dataset including around 1000 signs. We also have around 4000 videos including sign language sentences but not segmented in isolated signs. Additionally, it is possible to obtain several hours of sign language videos without any label. To generate the corpus the main tasks would be:
   a. Using the first 1000 signs as references to segment videos with sign sequences using the embeddings generated from CLIP. We can consider the similarity between video

sequences using frame embeddings or high-level characteristics using text and video frame embeddings.

b. Additionally, we will analyse the possibility to search for specific signs in unlabelled videos.

2. Extract 2D motion characteristics from videos. The skeleton description and its evolution can be extracted from selected or segmented videos using deep learning tools like OpenPose, Mediapipe, AlphaPose, etc. Several of these tools will be analysed and one will be selected to extract 2D motion information for the available videos.

Phase 2: Developing a deep learning algorithm able to associate high level sign characteristics to skeleton motion. In this phase, we plan the following activities:

1. The first task will be to evaluate different strategies to develop a motion generation system from sign characteristics. Several deep learning strategies will be implemented and evaluated: MotionCLIP, Transformers, VAE, etc.
2. Related to the DL architecture, an important aspect is how to code inputs and outputs. The most challenging aspect is how to represent the output. The output must be a sequence of motion information. This sequence can be modelled as a sequence of different states (using recurrent neural networks, for example) or as a whole description (global picture of the movement).

For the first year, the main target would be to address the first phase and analyse different strategies for the second phase.

SPEAKER DIARIZATION WITH MULTIMODAL INPUTS

Speaker diarization is the technical process of splitting up an audio recording stream that often includes several speakers into homogeneous segments. These segments are associated with each individual speaker. In short, this is what the “behind the scenes” process looks like when transcribing an audio recording file (like in Amazon Prime content). There exist several Artificial Intelligent techniques that solve the problem up to with 26 speakers in the same conversation with accurate results. But these results are normally over existing datasets, and the performance is reduced when you go to the real world in some concrete and more complex situations. Also, these techniques have been applied over the audio signal without the use of other information. All these solutions can be improved in different scenarios where other data is available such as video. The extraction of useful embeddings to identify speakers using both inputs, audio, and video, is something under exploration and some solutions have been presented in the literature.

The main target is to propose some use cases where these solutions can be applied using multi-modal information. The project can be divided in two main phases:

1. Identify use cases and collect appropriate datasets for the task. Pre-process the data and analyses techniques to extract previous information and convert the inputs to something appropriated for the artificial intelligent algorithms. This includes audio and video pre-processing, detection steps and tools for proper data manipulation.

2. Apply techniques for efficient embedding creation of multi modal data and create artificial intelligence solutions to solve the proposed problem. This involved, the creation of the networks, the training in custom datasets and the evaluation of the performance in the selected use cases.

POSE AND SPATIAL MOVEMENT AS INPUT FOR DYNAMIC CONTENT SEARCH & GENERATION

The idea behind this proposal is to explore the connection of poses and movements across the physical space to (video-audio) content search and/or generation, so movement and pose can serve as interactive input for this purpose. For example, the resulting system may be able to:

a) identify a yoga posture a user is performing and utilize that input to generate audio-visual feedback. Visual feedback may include other people performing the evolved posture (in case learning and training is the objective), suggestions on the next posture, or simply evocative/inspiring images to deal with the practice.

b) creatively respond to attract interest of a group or users sharing the same physical space performing different actions and paying attention to different spatial elements.

These are only examples of uses that may be instantiated in narrower applications focused on sports practice, training, marketing, assistance, etc.

In an explorative phase, the research line would include three main activities:

1) Posture & quality recognition component from live movement: in a first version, this component would be able to dynamically classify poses (e.g. yoga ones) using OpenPose (or similar tools) over real physical movement.

2) Translating pose & execution quality features to text and image features with content generation by using CLIP for image search.

3) Generating an integrated prototype combining both components on a camera-equipped space.

ENTANGLING AI-AUDIO SYNTHESIS MODELS AND MULTIMODAL REPRESENTATIONS

The world of cinema, performing arts, video games, and other mass-media industries are close to making a quantum leap in improving the quality of experience in sound and audio. Just as has happened for image and video in this last decade, AI technologies that made it increasingly difficult to distinguish an AI-generated video from a natural one must have a similar role in audio.

Audio poses significant challenges compared to image: the eye can be easily fooled, but the ear is more demanding. We want to evaluate how using multimodal correspondence between self-supervised representations, and AI-assisted synthetic audio generation techniques may help in this task.

This project seeks to take a step towards the controllable generation of audio effects in constrained visual environments, exploiting multimodal representations on WAV2CLIP pretrained representations and connecting those with an AI-generative audio effects synthesis. The goal is to enrich images (without audio) with sounds matching the scene.

We have already collected a database of footstep sounds for a limited number of conditions (certain types of soils and shoes). New databases will have to be found to complement this dataset. Moreover, GA recently developed an AI system based on a variational autoencoder for generating new samples of step sounds. We want to entangle the CLIP embeddings with the autoencoder to mix text with audio coherently. Moreover, we want to explore diffusion models as an alternative to the current VAE approach to improve control during audio generation.

Phase 1. Evaluation of the pre-trained net: WAV2CLIP.

Using the pre-trained model, we will start from audios available in our dataset and use WAV2CLIP to retrieve images that relate to these. From these images, we will retrieve audio relating to those images to test the variability on the WAV2CLIP.

Phase 2. Entangle audios with an audio effects generator

We will focus on the pre-trained network embedding to connect these with the current GAPS VAE model for synthesis, connecting the representations from CLIP and the VAE. We will also evaluate how these embeddings can be associated using diffusion models to combine a generative approach toward audio effects from CLIP multimodal embeddings.

Phase 3. Extension to short videos

We will extend this analysis to frames extracted from short videos to evaluate the generative capabilities of the model. We will use moving images (short clips) and look for the generated steps to match the moment in which the person executes the action. We will evaluate the possibility of performing this operation in real-time.

Phase 4. Extension for the sonorization of scenes in videogames

To incorporate generative audio, we will evaluate how embeddings and sonorization strategies can be applied to video games. We shall consider both offline (recorded scenes) and online.

For the first year, the main target would be to address the first phase and analyse different strategies for the second considering the VAE vs. diffusion model approach.

ZERO-SHOT SONORIZING OF VIDEO SEQUENCES

Storyboards provide visual representations of how a story will play out, scene by scene, event after event. After complete scripting, these storyboards are extensively used in movie, performance, and videogame industries, focusing on the main elements, viewpoints, and events. Score creators and sound designers use these to plan their work. Multimodal correspondence in the self-supervised image and audio representations may be used to build sound sequences that can successfully complement the video sequence.

This project will focus on short, simple video sequences to enrich them with audio content.

Phase 1. Evaluation of short sequences on the pre-trained net: AudioCLIP

First, we evaluate how the pre-trained AudioCLIP sequence sonorizes short video sequences displaying simple human motion. We will consider publicly available video datasets containing audio and video showing simple scenes in human-like motion, as similar as possible to those used to train AudioCLIP (e.g., UCF101, etc.), including real video and videogames.

Phase 2. Preparation of a scripted test set based on the pre-trained net: AvatarCLIP

We address the generation of short scripted (text) video sequences with simple avatars using the pre-trained AvatarCLIP. This we use to produce sequences displaying simple animated movements controlled through text. The generated scripted sequences will be tested on AudioCLIP to test whether the resulting multimodal embeddings are like those on the previous set.

We will initially focus on motion categories identified in the previous datasets and the footsteps effect. We already collected a database and developed a VAE on the latter.

Phase 3. Evaluate audio effects

We will evaluate the audio content and effects, focusing on the embeddings resulting from the AvatarCLIP and AudioCLIP.

If possible, we will consider evaluating learning strategies to adapt a model that can sonorize the images in a single shot, adapted from the publicly available datasets. If possible, at this point, we will connect Project 2 and Project 1.

Phase 4. Extension for the sonorization of scenes in videogames

We will evaluate how these embeddings and sonorization strategies can be applied to video games to incorporate generative audio, both offline (recorded scenes) and online.

For the first year, the main target would be to address the first phase and run preliminary tests on the embeddings corresponding to some basic scripted video sequences. These early tests shall help identify possible limitations prior to Phase 2.