Post-doctoral position

Research topic
Synthesis of videos driven by text and audio

Key words:
Video synthesis, generative adversarial networks, text to speech to video methods, image and video analysis and processing

Context of the study
Mon Petit Placement is a fintech startup created in 2017 in Lyon, whose mission is to democratize financial investment for all French and European investors.

To achieve this mission, Mon Petit Placement promotes proximity and trust between the Mon Petit Placement advisor and the customer who enters into a partnership with Mon Petit Placement through a 100% digital solution. To serve this proximity, Mon Petit Placement uses video to recreate a personalized digital link with the customer.

In this context, Mon Petit Placement seeks to automate the massive creation of videos in "talking head" format to accompany its customers throughout the lifetime of their investment, in a personalized manner. Therefore, the collaborative research project is set up between Mon Petit Placement and the LIRIS research laboratory. As part of this collaboration, we are looking for a post-doctoral researcher who will work in the field of text and audio driven video synthesis.

Description of the research project
In recent years, voice interaction with computers has made significant progress. Virtual agents offer a user-friendly human-machine interface while reducing maintenance costs. Speech-based interaction is already effective, as proved by Siri, Google Assistant or Alexa virtual agents, however, their visual counterpart is still far behind. The level of user engagement for audiovisual interactions is much higher than for purely audio interactions. Therefore, it is desirable to be able to associate face visual animations with the generated audio.

A notable advancement in video generation was made by a team at Stanford University in 2019 in partnership with Adobe [1]. Their work is aimed at enabling a video editing technology of a person's face-to-face scene to revise its speech script and adapt the rendering automatically based simply on this revised text.

The latest advances in the field of audio-driven face video synthesis were presented in [2]. The proposed approach generalizes across different people, to synthesize videos of a target actor with the voice of any actor from an unknown source or even synthetic voices that can be generated using standard text-to-speech approaches. Neural Voice Puppetry can generate videos with a visual synchronization quality superior to photorealistic audio and video reconstruction techniques.

During this post-doctoral contract, we want to work on the development of a prototype of a Text to Speech to Video technology with a sufficient level of accuracy for commercial use.
The final goal is the construction of a Text to Speech to Video technology allowing to generate in a totally automatic way, several minutes long video sequences of a person speaking in front of a camera (talking head) from a textual script. A generated video must have a sufficient quality to allow the illusion of an original video.

The generated videos must have the following characteristics:

1. The visual quality of the face should be photorealistic. It should not be distorted or develop erratic behaviors (inappropriate eye or head movements).
2. The visual quality of the agent's mouth and teeth have also to be photorealistic. This is currently still the major weakness of most video generation algorithms [3].
3. The synchronization of audio and video should be excellent.

In the first step, the research will focus on the generation of photorealistic videos of a person's face and mouth. In the second step, the development of the Text to Speech to Video solution will be proposed to allow the word change, word deletion, and the word addition not necessarily pronounced by the agent. And in the third step, we will work on the photorealism of the whole video by considering the other body movements tuned to the rhythm and the tone of the textual script (head movements, eyes movements, blinks, shoulders movements, etc.).

**Duration and place of work**

The funding covers 18 months of post-doc, the desired start is October 2021. The post-doctoral fellow will be attached to the LIRIS (Laboratory of Computer Science in Image and Information Systems) on the campus of the University Lyon 2 in Bron.

Some stages of work can be conducted in the office of Mon Petit Placement in Lyon.

**Supervisors**

The post-doc will be supervised by Iuliia Tkachenko and Serge Miguet (LIRIS). The project manager on the side of Mon Petit Placement is the technical director of the startup Thibault Jaillon.

**Successful candidate profile**

- The candidate must have a PhD in computer science, specializing in image and video processing
- Programming languages: Python/C++
- Neural network libraries: PyTorch/Keras/Tensorflow
- programming tools for image analysis: OpenCV
- Scientific knowledge: machine learning and deep learning, video analysis and processing
- Languages: French and/or English

**References**


Contact

Email:
iulii.tkachenko@univ-lyon2.fr
thibault@monpetitplacement.fr

Please provide a CV, a complete list of publications, a cover letter, two letters of recommendation or two reference names.