Offre de contrat post-doctoral

Sujet de recherche

Réseaux antagonistes génératifs pour la synthèse de vidéos pilotées par le texte et l'audio

Mots clés :

Synthèse de vidéos de visages, réseaux antagonistes génératifs, méthodes « *Text to Speech to Video* », analyse et traitement d'images et vidéos, permutation intelligente de visages

Contexte de l'étude

Au cours des dernières années, l'interaction vocale avec les ordinateurs a fait des progrès considérables. Les agents virtuels proposent une interface homme-machine conviviale tout en réduisant les coûts de maintenance. L'interaction basée sur la parole est déjà effective, comme le montrent les assistants virtuels Siri, Alexa ou Google Assistant, cependant, leur homologue visuel est encore très en retrait. Le niveau d'engagement des utilisateurs pour des interactions audiovisuelles est bien plus élevé que pour des interactions purement audio. Il est donc souhaitable de pouvoir associer des animations visuelles d'un visage à l'audio généré.
Un avancement notable dans la génération des vidéos est celui réalisé par les équipes de l'Université de Stanford en 2019 en partenariat avec Adobe [1]. Leurs travaux ont pour vocation de permettre l'exploitation d'une technologie d'édition vidéo d'une scène de personne face caméra afin de réviser son script de parole et d'adapter le rendu automatiquement sur la simple base de ce texte révisé.
Les dernières avancées dans le domaine de la synthèse de vidéos de visages audio-pilotées ont présentées dans [2]. L'approche proposée se généralise à travers différentes personnes, pour synthétiser des vidéos d'un acteur cible avec la voix de tout acteur de source inconnue ou même des voix synthétiques qui peuvent être générées en utilisant des approches standard de synthèse vocale. *Neural Voice Puppetry* permet de générer des vidéos d'une qualité de synchronisation visuelle supérieure aux techniques de reconstitution photo-réalistes audio et vidéo.
Pendant ce contrat post-doctoral, nous voulons travailler à la mise au point d'un prototype d'une technologie de *Text to Speech to Video* avec un niveau de précision suffisant pour un usage commercial.

Description du sujet de recherche

Le but final est la création d'une technologie de *Text to Speech to Video* permettant de générer de manière totalement automatique, des séquences vidéo de plusieurs minutes d'une personne parlant face caméra (*talking-head*) à partir d'un script textuel. Une vidéo générée doit avoir une qualité suffisante pour permettre l'illusion d'une vidéo originale.
Les vidéos générées doivent avoir les caractéristiques suivantes :

1) La qualité visuelle du visage doit être photo-réaliste. Il ne doit pas être déformé et ne pas développer de comportements erratiques (mouvements inadaptés des yeux ou de la tête).

2) La qualité visuelle de l'intérieur de la bouche et des dents de l'agent doit également être photo-réaliste. C'est actuellement encore la faiblesse majeure de la majorité des algorithmes de génération de vidéos [3].
3) La synchronisation de l'audio et de la vidéo doit être excellent.

Dans un premier temps, la recherche se focalisera sur la génération de vidéos photo-réalistes du visage et de la bouche d'une personne. Dans un deuxième temps, le développement de la solution *Text to Speech to Video* sera proposée pour permettre la prise en charge du déplacement de mots, de la suppression de mots, et de l'ajout de nouveaux mots non nécessairement prononcés par l'agent. Et dans un troisième temps, nous travaillerons sur le photoréalisme de la totalité de la vidéo en prenant en compte les autres mouvements corporels accordés au rythme et au ton du script textuel (mouvements de tête, mouvements des yeux, clignements des yeux, mouvements des épaules, etc.)

Durée et lieu de travail

Le financement couvre 18 mois de post-doc, le début souhaité est octobre 2020. Le post-doctorant sera attaché au LIRIS (Laboratoire d'Informatique en Image et Systèmes d'information) sur le campus de l'Université Lyon 2 à Bron.

Encadrement

Le post-doctorant sera suivi par Iuliia Tkachenko et Serge Miguet (LIRIS).

Profil recherché

- Le candidat doit avoir un doctorat en informatique, spécialisé dans le traitement des images et des vidéos
- Langages : Python/C++
- Bibliothèques de réseaux de neurones : PyTorch/Keras/Tensorflow
- Outils de programmation pour l'analyse d'image : OpenCV
- Connaissances scientifiques : apprentissage automatique et apprentissage profond, analyse et traitement des vidéos
- Langues : français et/ou anglais

Références

[1] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. Goldman, K. Genova, Z. Jin, C. Theobalt, M. Agrawala, "Text-based editing of talking-head video", ACM Transactions on Graphics (TOG), Vol.38, 2019.
[2] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, M. Niessner, "Neural Voice Puppetry: Audio-driven Facial Reenactment", ECCV 2020 (https://justusthies.github.io/posts/neural-voice-puppetry/)
[3] R. Tolosana, S. Romero-Tapiador, J. Fierrez, R. Vera-Rodriguez, "DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance", arXiv preprint arXiv:2004.07532, 2020.

Contact

Email : iuliia.tkachenko@univ-lyon2.fr

Merci de fournir un CV, une liste complète de publications, une lettre de motivation, deux lettres de recommandation.