ANNEXE 2
Présentation du projet de thèse

Résumé pour un public non scientifique (saisie libre de 1000 caractères max. impérativement en français)

Nous proposons d'associer les caractéristiques d'un mouvement humain avec ses paramètres (positions et déplacements des membres et du squelette) par une approche d'apprentissage (modèles issus de l'apprentissage profond par réseaux de neurones). A partir de descriptions de ces mouvements en langage naturel, nous cherchons à apprendre une correspondance entre les mouvements et leur "signification".
Un premier domaine d'application concerne l'analyse des mouvements d'un patineur ou d'une patineuse et l'association de ces mouvements avec le commentaire synchronisé du commentateur. Cela permettrait de générer des commentaires automatiquement, par exemple par détection des figures à partir de vidéos, ou de générer les mouvements à partir du commentaire (avatars).
Les techniques éprouvées dans cette première phase, pourraient être appliquées dans un contexte clinique. Ici, l'association entre des caractéristiques du mouvement de parkinsoniens pendant leur sommeil et leur description diagnostique permettrait de proposer un système d'aide au diagnostic qui pourrait alerter les praticiens en cas de signes précoces.

Présentation scientifique du projet de recherche (objectifs, méthode, …) (saisie libre de 5000 caractères max.) :

Le projet SemTaxM (Sémantique et Taxonomie du Mouvement), projet de recherche transverse mené par des chercheurs des centres de recherche LGI2P/IMT Mines Alès et EuroMov a pour objectif d'étudier les théories dédiées à la classification du mouvement (au sens de la classification des objets d'études), afin de faire émerger une théorie linguistique du mouvement. Nous envisageons d'identifier empiriquement des invariants dans les paramètres du mouvement (composants, position, direction…) par l'apprentissage de représentations multimodales du mouvement (croisement d'estimation de poses à partir de vidéo avec des modèles de langues neuronaux). Ce sujet de thèse, « Apprentissage de représentations multimodales pour *sémantiser* les paramètres du mouvement » constitue un premier pas vers cet objectif en explorant les techniques d'apprentissage de représentations multimodales pour la *sémantisation* de poses issues du mouvement (vidéos).

Peng et al. (2019) ont proposé un système pour l'apprentissage de l'imitation du mouvement humain au travers de l'apprentissage par renforcement dans la prestigieuse conférence SIGGRAPH. Le système permet à un personnage virtuel d'apprendre à imiter le mouvement d'un modèle humain dans une vidéo, cela permet la génération automatique de mouvements réalistes, sans avoir besoin d'utiliser un onéreux système de motion capture, mais mène aussi à l'obtention d'un modèle de représentation générique capturant les propretés intrinsèques du mouvement. Nous souhaitons exploiter un principe similaire, mais en y ajoutant une dimension multimodale qui viserait à apprendre de manière jointe un modèle du mouvement (estimation de pose 3D par apprentissage profond couplée à un modèle d'apprentissage de représentations permettant l'imitation) avec un modèle de langage, en particulier les modèles de langage contextualisés profonds qui ont révolutionnés le domaine du Traitement Automatique du Langage Naturel.
Cette direction de recherche est d'actualité si l'on se réfère à l'apparition de travaux allant dans le sens de cette combinaison multimodale des paramètres du mouvement et de descriptions sémantiques explicites. En effet, Ahuja et Morency (2019) ont proposé un système qui permet d'apprendre une

représentation par plongement jointe, combinant un corpus de poses (poses successives d'un squelette dans le temps) et un modèle de plongements sémantiques classiques (Word2Vec, Mokolov et al. 2012). Cependant, dans cette approche nous constatons l'utilisation de techniques classiques ne faisant pas usage des dernières avancées à base d'apprentissage profond, à la fois du côté mouvement (Estimation de pose 3D, par ex. VideoPose3D) et du côté du modèle de langue utilisé (i.e. modèles de langue contextualisés profonds). L'espace de plongement ainsi appris sera donc limité : notamment impossibilité de capturer des phénomènes multiculturels complexes (compositionnalité du langage et du mouvement, invariants), qui sont pourtant des propriétés indispensables dans le contexte de l'utilisation de tels systèmes pour l'étude des caractéristiques du mouvement.
Ainsi nous souhaitons développer une nouvelle famille d'architectures d'apprentissage de représentation profonde multimodale de bout-en-bout (end-to-end), partant d'un côté directement de vidéos par l'utilisation du système VideoPose3D (Pavllo et al. 2019) en tant que composant de l'architecture, puis en utilisant du côté langage des modèles de langues contextualisés profonds (architecture à base de *transformeurs* permettant à la fois la discrimination et la génération, citer GPT-2).
L'entraînement se fera en suivant les paradigme d'entraînement multitâche par transfert sur la base de jeux de données généraux (domaine public, par exemple Shruti et al. 2019, Shi et al. 2019, Wang et al. 2019) avec ensuite un raffinement des représentations pour correspondre à des cas applicatifs différents. L'évaluation pourra se faire sur l'ensemble de ces jeux de données, par rapport à un système de référence qui sera celui proposé par (Ahuja et Morency 2019).
Le cas applicatif envisagé est celui de l'aide au diagnostic de Parkinsoniens sur la base de vidéos de mouvements lors du sommeil (collections de vidéos de grande taille mis à disposition par la Clinique Beau Soleil). Les mouvements des Parkinsoniens lors du sommeil sont souvent révélateurs de la progression de la maladie, ainsi ces enregistrements constituent un outil diagnostique standard. Actuellement le visionnage des vidéos est fait manuellement par les cliniciens, qui décrivent les anomalies observées tout le long de la vidéo. Ce diagnostic est très chronophage et donc coûteux. Il est donc souhaitable de pouvoir aider les cliniciens à repérer les moments saillants des vidéos pour réduire leur temps passé à ces analyses.
Il est visé un article de rang A+ par an: 1/Article de revue état de l'art ; 2/Article de conf. ; 3/Article de conf. sur le cas d'application sur le diagnostic de Parkinsoniens.

Références

C. Ahuja and L. Morency, "Language2Pose: Natural Language Grounded Pose Forecasting," *2019 International Conference on 3D Vision (3DV)*, Québec City, QC, Canada, 2019, pp. 719-728. doi: 10.1109/3DV.2019.00084

Distributed representations of words and phrases and their compositionality
T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean - Advances in neural information processing systems, 2013

Palaskar, Shruti, Jindrich Libovický, Spandana Gella, and Florian Metze. "Multimodal Abstractive Summarization for How2 Videos." arXiv preprint arXiv:1906.07901 (2019).

Peng, Xue Bin, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. "Sfv: Reinforcement learning of physical skills from videos." ACM Transactions on Graphics (TOG) 37, no. 6 (2019): 178.

Pavllo, Dario, Christoph Feichtenhofer, David Grangier, and Michael Auli. "3D human pose estimation in video with temporal convolutions and semi-supervised training." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7753-7762. 2019.

Yang, Xitong, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A. Bernal, and Jiebo Luo. "Deep multimodal representation learning from temporal data." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5447-5455. 2017.

Shi, Botian, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. "Dense procedure captioning in narrated instructional videos." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6382-6391. 2019.

Wang, Weiying, Yongcheng Wang, Shizhe Chen, and Qin Jin. "YouMakeup: A Large-Scale Domain-Specific Multimodal Dataset for Fine-Grained Semantic Comprehension." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5136-5146. 2019.

Liens du sujet de thèse avec les activités de l'unité de recherche d'accueil (saisie libre de 1000 caractères max.) : Dans la continuité des collaborations et actions déjà engagées entre les deux institutions, une réflexion est menée sur la création d'une unité mixte de recherche entre le LGI2P de IMT Mines Alès et EuroMov de l'Université de Montpellier, « Santé Numérique en Mouvement ». Cette thèse s'intègre dans cette perspective. Le sujet « Apprentissage de représentations multimodales pour *sémantiser* les paramètres du mouvement » se positionne sur l'Axe en devenir de la nouvelle UMR SNM Sémantique et Taxonomie du Mouvement (SemTaxM, future équipe de recherche pressentie à l'issue du quinquennal) donc les deux porteurs sont Andon Tchechmedjiev (Traitement Automatique du Langage Naturel, encadrant de proximité) et Julien Lagarde (Taxonomie et Paramétrique du Mouvement, Co-encadrant EuroMov). Cet axe représente un sujet d'étude interdisciplinaire sur les trois versant de la future UMR SNM (IA, Mouvement, Santé). La directrice de thèse (S. Ranwez, Ingénierie des Connaissances) est aussi dans SemTaxM.