Subject: Self-supervised learning for automatic gesture recognition.

Supervisors: Assoc. Prof. Decky Aspandi LATIF, Prof. Titus ZAHARIA
Duration: 6 months

Background:

Gesture is an important nonverbal communication that facilitates many human interactions in many settings, including driving or for communication of disabled individual [1]. As such, acquiring capability to automatically recognized this means of communication is quite substantial to enable better human and computer interactions. Current progress in developing automatic human gesture recognition has been accelerating and currently moving toward the use of several modalities to leverage in more information to achieve better accuracy [2]. However, this progress is hampered the expensive human annotation required to establish the necessary dataset [3]. To tackle this, one solution is to use self-supervised learning that allows for the expansions of the learning process applied to unlabeled data. This is achieved through the use of pre-text task (e.g. basic image transformation) to learn data characteristics [4,5], which subsequently be transferred forward to the end task, i.e. gesture recognition. This method has been largely applied in several other computer vision tasks [6], however its application to gesture recognition, particularly in multi-modal setting is still limited [7].

Target:

Given current limitations, this M1 internship aims to apply self-supervised learning to uni and possibly multi-modal inputs for automatic gesture recognition. Furthermore, ablation study to evaluate the impact of each elements of the method (self-supervised learning and multi-modal input) with respect to the achieved results will be performed. Finally, reaching (or at least intent to achieve) state of the art accuracy in comparison to other alternatives in the literature also constitutes to a relevant objective.

Methodology:

The first step of this internship will involve the use basic of self-supervised learning [4] with the aims of learning an image or depth characteristics in uni-modal setting using pre-text learning. Then using both modality inputs, the embedding technique can be applied to enable closed representations between these modalities [5,7]. Between these settings, the methods results should be observed to evaluate the efficacy of the introduced technique. In addition, several other technique, such as other fusion mechanism [8,9] or other self-supervised learning may also be explored to advance the developed technique further [3]

Dataset and evaluations:

Two datasets can initially be used for the evaluations of the developed method: SkiG [10] and IsoGD [11]. Both of these datasets contain continuous recording of RGB and Depth modality for gesture recognition task, that allows for modality learning. As the learning progress, other datasets such as chalearn [12] may be utilised to include other modality, such as audio. As for quantitative evaluations, the reported scores in the literature can be used for relative comparisons, with (harmonic) accuracy (in percentage unit) and F1 score can be used [1].

Expected results:

1. Evaluations and analysis of the impact of the self-supervised learning to each and both modality in terms of accuracy gained.
2. Further observations of the influence of modality fusion mechanism in relation to the first expected results.
3. Reported comparison results from the proposed method with other methods in literature. Ideally, the proposed method should at least produce competitive results.

Timeline and Research Lab

1. The research is expected to be conducted for six months with the results to be communicated in the Master Thesis and possibly conference publication.
2. The research will be carried out within the ARTEMIS department of Télécom SudParis, ARMEDIA team and SAMOVAR laboratory, in conjunction with France Television for study cases.

Requirements

1. Knowledge in Machine Learning.
2. Knowledge in Deep Learning.
3. Knowledge of Image Processing.

References:

[1] Mohamed, Noraini, Mumtaz Begum Mustafa, and Nazean Jomhari. "A review of the hand gesture recognition system: Current progress and future directions." IEEE Access 9 (2021): 157422-157436.
[2] Zhu, Guangming, et al. "Multimodal gesture recognition using 3-D convolution and convolutional LSTM." Ieee Access 5 (2017): 4517-4524.
[3] Zong, Yongshuo, Oisin Mac Aodha, and Timothy Hospedales. "Self-Supervised Multimodal Learning: A Survey." arXiv preprint arXiv:2304.01008 (2023).
[4] Schneider, T., Qiu, C., Kloft, M., Latif, D. A., Staab, S., Mandt, S., & Rudolph, M. (2022). Detecting anomalies within time series using local neural transformations. arXiv preprint arXiv:2202.03944.
[5] Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." Proceedings of the IEEE international conference on computer vision. 2017. - image and audio.
[6] Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, *9*(1), 2.
[7] Tamhane, Aniruddha, Jie Ying Wu, and Mathias Unberath. "Multimodal and self-supervised representation learning for automatic gesture recognition in surgical robotics." arXiv preprint arXiv:2011.00168 (2020).
[8] Roitberg, Alina, et al. "Analysis of deep fusion strategies for multi-modal gesture recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.
[9] Ouenniche, Kaouther, Ruxandra Tapu, and Titus Zaharia. "Vision-text cross-modal fusion for accurate video captioning." IEEE Access (2023).
[10] Learning discriminative representations from RGB-D video data
[11] Wan, Jun, et al. "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2016.
[12] Escalera, Sergio, et al. "Multi-modal gesture recognition challenge 2013: Dataset and results." Proceedings of the 15th ACM on International conference on multimodal interaction. 2013.