LIMOS – Axe SIC

Directeur de thèse : Vincent Barra (vincent.barra@uca.fr)
Collaboration : Freddy Maso, Directeur de la performance et de l'innovation, ASM Omnisports

Title of PhD subject. Behavioral Analysis of Videos Using Multimodal Language Models

Summary :

AI-based sports video analysis leverages machine learning and computer vision to transform raw sequences into actionable insights for athletes, coaches, and broadcasters.

The objective of this thesis is to develop innovative methods for behavioral analysis of elite athletes from videos recorded in real-world conditions. The research avenues considered involve the development of models for:

- Behavioral analysis based on videos, in the presence of constraints inherent to recording conditions and the specificities of the sport studied,
- Multimodal language models, adapted to the study context, and more particularly a foundation model that could then be fine-tuned.

This topic proposes to address four issues, allowing to overcome two scientific challenges and answer two application questions:

1- How to adapt a video foundation model (ViFM) to the specific problem. While some models exist (e.g. ActivityNet [1]), they cannot in their current state meet the challenges and constraints of the field addressed.
2- How to build a multimodal language model (MLM) that allows efficient querying of video sequences from prompts? Currently, the retrieval task in videos (e.g. T2V [2]) does not, to our knowledge, use MLM, and therefore does not exploit the potential of these algorithms that have proven their effectiveness and performance in many domains.
3- How can these models help automate the analysis of play produced by both teams?
4- How, through indicators, can these models analyze player performance with progression and recruitment objectives?

The models developed in this work will be fed by numerous data provided by the partner, for which ground truth is available. The partner's application expertise will additionally allow for a posteriori validation of the results produced by the algorithms.

While the main application target is the study of sports videos, the algorithms produced during this thesis can be adapted through fine-tuning to many other problems (attack detection in ATMs, fall detection, ...).

[1] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.
[2] Q. Ye, G. Xu, M. Yan, H. Xu, Q. Qian, J. Zhang, and F. Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 15405–15416, October 2023.