Title: Human behavior understanding in videos using multimodal foundation models

Type: Research internship in Computer Vision

Level: Master 1 or 2, Fourth or Fifth-year Engineer in Computer Science or close-related field.

Supervision: Carlos Crispim Junior

Location:

LIRIS UMR CNRS 5205,
Université Lumière Lyon 2,
5 avenue Pierre Mendès-France, 69676 Bron CEDEX

Keywords: deep learning, video understanding, human behavior analysis

Context of the study:

Human behavior understanding is a key task for several fields of application, from human assisted living and disease diagnosis in healthcare to industry problems, like task training and completion evaluation. Deep neural networks and, more recently, multimodal foundation models have brought a new level of performance to research problems in video understanding (*e.g.,* Dino v3, VideoLLaMA, InternVideo2). However, the performance of such methods in behavior understanding, like emotion recognition, is still limited compared to generic scene understand*ing (*Lian et al., 2024*).*

This internship subject will study and evaluate the latest multimodal foundations models as building blocks for a pipeline for human behavior understanding. We will focus on methods capable of describing emotion and gesture recognition in long videos and explore their performance outside datasets with controlled conditions (*i.e.*, in the wild).

Tasks:

- Revise the state of the art on methods for multimodal video understanding applicable for behavior understanding, identifying their limitations on the characterization of the target behavioral aspects.
- Propose a spatio-temporal based deep neural pipeline that can detect the target behavioral events in space and time.
- Write a research article to share the developed work with the computer vision community, accompanied by an open-source repository to foster reproducible research.

Profile of the candidate:

We are looking for a motivated candidate with a strong background in computer science or applied mathematics.

- The candidate must currently be enrolled in a Master 1 or 2 program, or be in the final years of engineering school (Bac+4 or +5 in France)
- Experience in image processing, computer vision, and/or machine learning will be a plus.

If the internship leads to an international publication, we may study opportunities to pursue the research carried out with a PhD in a similar topic.

Language: French or English

EXPECTED SKILLS:

- Mastering of Python language
- OpenCV library
- Versioning tools (GIT)

The following skills would be considered as a plus:

- Framework PyTorch or TensorFlow.
- Docker-like tools and platforms

DURATION: 4-6 months

EXPECTED INTERNSHIP PERIOD: Late April-October, with an imposed summer break

SALARY: “gratification de stage” in France

CONTACT : carlos.crispim-junior@liris.cnrs.fr

APPLICATION: Curriculum Vitae and last two-year university transcripts, and recommendation Letter

RELATED BIBLIOGRAPHIC REFERENCES:

- Zheng Lian, *et al.*, GPT-4V with emotion: A zero-shot benchmark for Generalized Emotion Recognition, Information Fusion, Volume 108, 2024, 102367, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2024.102367.
- Boqiang Zhang, *et al.*, VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding, 2025, https://arxiv.org/abs/2501.13106
- Yi Wang, *et al.* InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. In Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXV. Springer-Verlag, Berlin, Heidelberg, 396–416. https://doi.org/10.1007/978-3-031-73013-9_23
- Oriane Siméoni, *et al.*, DinoV3, 2025, https://arxiv.org/abs/2508.10104