3iA Côte d'Azur
Interdisciplinary Institute
for Artificial Intelligence

Proposal for a PhD thesis

- Research axis of the 3IA: Axis 2: AI for Integrative Computational Medicine
- Supervisor: Francois Bremond
- Research group: STARS team at INRIA Sophia Antipolis

I. Title

Semantic Modelling within Transformers for Action Detection in Untrimmed Videos

II. General objective

Action detection is a challenging computer vision problem which targets at finding precise temporal boundaries of actions occurring in an untrimmed video. Many studies on action detection focus on videos with sparse and well-separated instances of action. For instance, action detection algorithms on popular datasets like THUMOS and ActivityNet generally learn representations for single actions in a video. However, in daily life, human actions are continuous and can be very dense. Every minute is filled with potential actions to be detected and labelled. The methods designed for sparsely labelled datasets are hard to generalize to such real-world scenarios.
Towards this research direction, several methods [2, 3, 5] have been proposed to model complex temporal relationships and to process datasets like Charades, TSU [1] and MultiTHUMOS. Those datasets encompassing real-world challenges share the following characteristics: Firstly, the actions are densely labelled and background instances are rare in these videos compared to sparsely labelled datasets. Secondly, the video has rich temporal structure and a set of actions occurring together often follows a well-defined temporal pattern. For example, drinking from bottle always happens after taking a bottle and reading a book also related to opening a book. Moreover, humans are great at multitasking, multiple actions can co-occur at the same time. For example, reading book while drinking water.

So, the main question is how semantics can be modelled to help recognizing complex temporal relationships in videos?
Typical situations that we would like to monitor are Eating and Drinking (how much? how often?) or Cooking (detect behavior that might lead to dangerous situations or non-completion of the task).
The system we want to develop will help senior people and their relatives to feel more comfortable at their home, since scene understanding intends to help at recognizing potentially dangerous situations and reporting to caregivers if necessary.

IV. PhD objective

In this work we would like to go beyond Deep Learning by incorporating some semantic modelling within the Deep Learning pipeline, which consists of a combination of CNN and transformers [5] to be able to model the complex action patterns in untrimmed videos. These complex action patterns include composite actions and concurrent actions existing in long untrimmed videos.

Existing methods [3, 5, 6] have mostly focused on modelling the variation of visual cues across time locally or globally within a video. However, these methods consider the temporal information without any further semantics. Real-world videos contain many complex actions with inherent relationships between action classes at the same time steps or across distant time steps. Modelling such class-temporal relationships can be extremely useful for locating actions in those videos.

In this work, we focus on semantic modelling for improving action detection performance. Videos may contain rich semantic information such as objects, actions, and scenes. Relationships among different semantics are high-level knowledge which is critical for understanding the video content. Therefore, semantic relational reasoning can help determine the action instance occurrences and locate the actions in the video, especially for complex actions in video. For handling these challenges, Class-Temporal Relational Network (CTRN) [4] has been proposed to explore both the class and temporal relations of detected actions.
To go beyond the above, a first attempt may consist to:
(1) Effectively extracting action-relevant semantics from real-world untrimmed videos.
(2) Modelling the cross semantic relations to enhance the action detection performance.
(3) Incorporate the modelled semantics within the Deep Learning pipeline.

To extract the relevant semantics, large Language-Vision model could be used.

This work will be conducted within the Cobtek team from Nice Hospital, who is specialized in clinical trials for older adults with dementia.
The evaluation of proposed frameworks and models should be performed on public datasets which contains everyday activities like Charades, MultiTHUMOS, and homecare datasets like TSU [1].

There is a possibility of conducting first an internship, before the PhD thesis.

IV. Prerequisites

Strong background in C++/Python programming languages,
Knowledge on the following topics is needed:

Machine learning,
Deep Neural Networks frameworks,
Probabilistic Graphical Models,
Computer Vision, and
Optimization techniques (Stochastic gradient descent, Message-passing).

V. Calendar

1st year:

Study the limitations of existing activity recognition algorithms.
Depending on the targeted activities, data collection might need to be carried out.
Propose an original algorithm that addresses current limitations on inference.
Evaluate the proposed algorithm on benchmarking datasets,
Write a paper

2nd year:

Investigation of feasibility/appropriateness of the framework in practical situations
Propose an algorithm to address model learning task in supervised settings
Write a paper

3rd year:

Optimize proposed algorithm for real-world scenarios.
Write a paper, and the PhD Manuscript

VI. Bibliography:

[1] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. Transactions on Pattern Analysis and Machine Intelligence, TPAMI, ISSN: 0162-8828, Digital Object Identifier: 10.1109/TPAMI.2022.3169976, PAMI 2022.

[2] Rui Dai, Srijan Das, and Francois Bremond. Learning an augmented RGB representation with crossmodal knowledge distillation for action detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13053–13064, October 2021.

[3] Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. PDAN: Pyramid dilated attention network for action detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2970–2979, January 2021.

[4] R. Dai, S. Das and F. Bremond. CTRN: Class Temporal Relational Network For Action Detection. In Proceedings of the 32nd British Machine Vision Conference, BMVC 2021, hal-03383140v2, United Kingdom, Virtual, November 22-25, 2021.

[5] R. Dai, S. Das, K. Kahatapitiya, M. Ryoo and F. Bremond. MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, Hybrid, June 19-23, 2022.

[6] AJ Piergiovanni and Michael S Ryoo. Temporal gaussian mixture layer for videos. International Conference on Machine Learning (ICML), 2019.

VIII. Contact:

Francois.Bremond@inria.fr

Apply by sending an email directly to the supervisor.

The application will include:

- Letter of recommendation of the supervisor indicated above
- Curriculum vitæ.
- Motivation Letter.
- Academic transcripts of a master’s degree(s) or equivalent.
- At least, one letter of recommendation.
- Internship report, if possible.

All the requested documents must be gathered and concatenated in a single PDF file named in the following format: LAST NAME of the candidate_Last Name of the supervisor_2023.pdf