Temporal action detection for untrimmed videos based on deep neural networks

A propos du centre ou de la direction fonctionnelle

The Inria Sophia Antipolis - Méditerranée center counts 37 research teams and 9 support departments. The center's staff (about 400 people including 100 Inria employees) is composed of scientists of different nationalities (250 Foreigners of 50 nationalities), engineers, technicians and administrators. 1/3 of the staff are civil servants, the others are contractual. The majority of the research teams at the center are located in Sophia Antipolis and Nice in the Alpes-Maritimes. Six teams are based in Montpellier and a team is hosted by the computer science department of the University of Bologna in Italy. The Center is a member of the University and Institution Community (ComUE) "Université Côte d'Azur (UCA)".

Contexte et atouts du poste

To support this work, we have a full team of researchers specialized in human behaviors [15, 11, 3, 2], from experts in activity recognition, people detection and tracking, machine learning, up to medical doctors specialized in behavioral disorders. The STARS team has been working on analytics video untrimmed video since 2014. The "Understanding Platform" system developed in STARS, detects mobile objects, tracks their trajectory and recognizes related behaviors predefined by experts. This platform contains several techniques for the detection of people and the recognition of human postures and gestures from one conventional camera. We have access to large cohorts of patients and can collect video datasets, dedicated to behavioral disorders, such as the ones induced by dementia. We also have large storage resources and a heavy GPU farm, from which 28 GPU nodes are dedicated to STARS team.

Mission confiée

Temporal action detection in untrimmed videos (long video containing several actions) is an important task for monitoring patients, building robots for assisting and other healthcare applications. Although several approaches, including the Deep Convolutional Networks (CNNs), have significantly improved performance on action classification, they still struggle to achieve precise temporal action localization in untrimmed videos. Temporal action detection aims at not only recognizing the action category but also detecting the beginning and ending of an action instance. Most temporal action detection frameworks consist of two parts: action boundary proposition and action classification.

The first task, "action boundary proposal", consists in determining the temporal boundaries of each action instance. Existing work as [10, 14, 13, 8, 6] have low precision on this detection of temporal boundaries. These algorithms meet difficulties for detecting long complex actions (e.g. cooking). Besides, they usually fail to detect the actions where the duration varies significantly, from a couple of seconds to few minutes. On the other hand, to obtain high localization accuracy, a large number of window scales and small sliding steps would be needed, which can lead to dramatically increased computational cost. Hence, we lack of an efficient and robust algorithm for localizing the actions.

The second task is "action classification" which is to classify accurately a video with action labels. Recently we have designed high performing model [1, 2], which can get more than 90% accuracy on several public datasets as NTU-RGB+D [9]. However these models fail to achieve high performance in real life settings datasets. Errors come with handling real life challenges, such as high environment diversity, multiple-view settings, low awareness of camera, high duration variation, etc. In addition, long action recognition with complex actions (e.g. making coffee in a coffee machine) and fine-grained actions with different objects (e.g. drinking from a cup or a bottle) are still unsolved tasks. Hence, we still need robust algorithms for action classification in real life settings.

The algorithm that we want to develop will be deployed in real life settings, to help senior people and their relatives to feel safer at home since video analytics intends to detect potentially dangerous situations and to report critical situations to caregivers.

In this PhD work, we would like to go beyond Deep Learning by taking advantage of CNN based network for action classification and embedded them into a temporal action detection framework for action localization to address complex human daily living datasets.

The challenge is to design a method that can process an untrimmed video in both online and offline manner and so to detect automatically the beginning and the ending of the targeted actions. A typical system can include 2 sub-networks: generating temporal proposals and classifying proposed candidates. The former is to produce a set of class-agnostic temporal regions that potentially reflect actions of interest, while the latter is to determine whether each candidate actually corresponds to an action and what class it belongs to. CNNs, RNN could be used in this system.

The evaluation of proposed frameworks and models should be performed on public live videos and datasets which contain daily activities like A4A [3], THUMOS [4], PKU-MMR [7], OHULIA [32] and Smarthome.


Principales activités

Calendar:

1st year: Study the limitations of existing activity recognition and temporal detection 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 situation. Propose an algorithm to address model learning task in semi-supervised settings, write a paper and write PhD manuscript.

3rd year: Optimize proposed algorithm for real-world scenarios. Write a paper and PhD Manuscript.

Prerequisites:

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

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

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Rémunération

Duration: 36 months
Location: Sophia Antipolis, France
Gross Salary per month: 1982 € brut per month (year 1 & 2) and 2085€ brut/month (year 3)