HUMAN ACTION RECOGNITION IN VIDEOS WITH LOCAL REPRESENTATION

by
Michal Koperski
Supervisor: Francois Bremond
STARS, Inria Sophia Antipolis, France

ABSTRACT

This thesis targets recognition of human actions in videos. Action recognition can be defined as the ability to determine whether a given action occurs in the video. This problem is complicated due to the high complexity of human actions such as appearance variation, motion pattern variation, occlusions, etc.

Recent advancements in either hand-crafted or deep-learning methods significantly improved action recognition accuracy. But there are many open questions, which keep action recognition task far from being solved.

Current state-of-the-art methods achieved satisfactory results mostly base on features, which focus on a local spatio-temporal neighborhood. But human actions are complex, thus the following question that should be answered is how to model a relationship between local features, especially in spatio-temporal context. In this thesis, we propose 2 methods which try to answer that challenging problem. In the first method, we propose to measure a pairwise relationship between features with Brownian Covariance. In the second method, we propose to model spatial-layout of features *w.r.t.* person bounding box, achieving better or similar results as skeleton based methods. Our methods are generic and can improve both hand-crafted and deep-learning based methods.

Another open question is whether 3D information can improve action recognition. Currently, most of the state-of-the-art methods work on RGB data, which is missing 3D information. In addition, many methods use 3D information only to obtain body joints, which is still challenging to obtain. In this thesis, we show that 3D information can be used not only for joints detection. We propose a novel descriptor which introduces 3D trajectories computed on RGB-D information.

Finally, we claim that ability to process a video in real-time will be a key factor in future action recognition applications. All methods proposed in this thesis are ready to work in real-time. We proved our claim empirically by building a real-time action detection system, which was successfully adapted by Toyota company in their robotic systems.

In the evaluation part, we focus particularly on daily living actions – performed by people in their daily self-care routine. In the scope of our interest are actions like eating, drinking, cooking. Recognition of such actions is particularly important for patient monitoring systems in hospitals and nursing homes. Daily living action recognition is also a key component of assistive robots.

To evaluate the methods proposed in this thesis we created a large-scale dataset, which consists of 160 hours of video footage of 20 senior people. The videos were recorded in 3 different rooms by 7 RGB-D sensors. We have annotated the videos with 28 action classes. The actions in the dataset are performed in un-acted and unsupervised way, thus the dataset introduces real-world challenges, absent in many public datasets.

Finally, we have also evaluated our methods on publicly available datasets: CAD60, CAD120 and MSRDailyActivity3D. Our experiments show that the methods proposed in this thesis improve state-of-the-art results.