PhD in Computer and Control Engineering

Research Title: First person action recognition from multi-modal data

Funded by	Comitato ICT

Supervisor	Barbara Caputo barbara.caputo@polito.it Marco Mezzalama marco.mezzalama@polito.it

Contact	https://scholar.google.it/citations?hl=en&user=mHbdIAwAAAAJ&view_op=list_works

Context of the research activity	While the vast majority of existing digital visual data has been acquired from a third person perspective, egocentric vision is soon going to become a key technology for supporting new developments in assistive driving, enhanced cognition and many more applications. This in turn will require the development of visual recognition algorithms able to deal with the challenges of this new scenario, from transfer learning across different actors to action anticipation, and so forth. This PhD thesis will explore first person action recognition on RGB, RGB-D and 3D point cloud data, i.e. when not only standard images are available, but also 3D and/depth information. The thesis will develop new algorithms for action recognition that can take advantage of the various modalities when they are present, and that at the same time make it possible to hallucinate them from reference benchmark when they are not, in order to simulate a modal completion of data and in crease the robustness of the system. The developed algorithms will be tested on various applications, from human-robot interaction, to industrial robot applications and assistive driving.

Objectives	Automated analysis of videos for content understanding is one of the most challenging and well researched areas in computer vision and multimedia and possesses a vast array of applications ranging from surveillance, behavior understanding, video indexing and retrieval, human-machine interaction, etc. The majority of researchers working on video understanding problem focuses on action recognition from distant or third person views while egocentric activity analysis has been investigated more recently. Action recognition involves identifying a generalized motion pattern of hands such as take, put, stir, pour, etc. whereas activity recognition concerns more fine-grained composite patterns such as take bread, take water, put sugar in coffee, put bread in plate, etc. For developing a system capable of recognizing activities, it is pertinent to identify both the hand motion patterns as well as the objects on to which a manipulation is being applied to. Although commonly forgotten, 2.5D (depth) and/or 3D information should be taken into consideration, both for better understanding actions and their context, and also for making it possible to translate results in this field into real life scenarios such as human-robot interaction, personal logging and assisted driving in a straightforward manner. Open challenges are how to leverage effectively over these modalities without a memory and/or computational explosion, and how to develop algorithms flexible enough to move from one device, equipped with all sensors, to another with an impoverished setting. This can be seen as a domain adaptation problem across modalities. This PhD thesis will tackle these problems, extending previous state of the art first person action recognition architecture to deal also with 2.5D and 3D point cloud data and to deal with domain adaptation across different acquisition devices. Specifically, we will tackle these issues by learning to hallucinate sensor modalities when available, and leveraging over such hallucinators when deployed on impoverished platforms. To do so, we will borrow from previous work on RGB-D object categorization [1deco] and domain generalization [2adage]. Progress will be assessed on existing public benchmarks, as well as data collected by the PhD student using a Pupil wearable helmet, in driving and assistive robotic scenarios.

	[1] F. M. Carlucci, P. Russo, B. Caputo. DECO: deep depth colorization. IEEE Robots and Automation Letters, 2018. [2] F. M. Carlucci, P. Russo, T. Tommasi, B. Caputo. Agnostic Domain Generalization. arXiv preprint arXiv:1808:01102.

Skills and competencies for the development of the activity	The candidate is expected to have a Master of Science degree in Computer Engineering, Computer Science or similar fields, and prior knowledge in machine/deep learning, computer vision and statistical methods. Projects, tesi or publications in these areas will constitute a plus. The candidate is also expected to have programming experience in python