Robot Learning Workshop

October 14-15, 2019

Yiannis Aloimonos
University of Maryland

Show and tell: Robots Learning Actions from Vision and Language

Context-free grammars have been in fashion in linguistics because they provide a simple and precise mechanism for describing the methods by which phrases in some natural language are built from smaller blocks. Also, the basic recursive structure of natural languages, the way in which clauses nest inside other clauses, and the way in which lists of adjectives and adverbs are followed by nouns and verbs, is described exactly. Similarly, for manipulation actions, every complex activity is built from smaller blocks involving hands and their movements, as well as objects, tools and the monitoring of their state. Thus, interpreting a "seen" action is like understanding language, and executing an action from knowledge in memory is like producing language. Several experiments will be shown interpreting human actions in the arts and crafts or assembly domain, through a parsing of the visual input, on the basis of the manipulation grammar. This parsing, in order to be realized, requires a network of visual processes that attend to objects and tools, segment them and recognize them, track the moving objects and hands, and monitor the state of objects to calculate goal completion. These processes will also be explained and we will conclude with demonstrations of robots learning how to perform tasks by watching videos of relevant human activities.