Keynote

People in Motion: Pose, Action, and Communication

Stan Sclaroff

Boston University

This talk will give an overview of some of the research in the Image and Video Computing Group at Boston University related to tracking, analysis, recognition and retrieval of images and video based on humans and their actions.

First, efficient methods for inference of human pose will be presented, where a tree-based articulated pose model incorporates higher-order constraints. In one approach, scale and rotation invariant matching is made tractable through the use of linearly augmented trees; this enables efficient optimization over continuous scale and rotation parameters. Our experiments on ground truth data and a variety of real images and videos show that the proposed method is ef?cient, accurate and reliable. In another approach, articulated pose estimation with loopy graph models is made efficient via a branch-and-bound strategy for finding the globally optimal pose; the algorithm converges rapidly in practice due to a novel method for quickly computing tree-based lower bounds.

Second, methods for learning human action models from Web images and video will be presented. The methods are unsupervised in the sense that they require no human intervention other than the action keywords to be used to form text queries to Web image and video search engines. Thus, it is easy extend the vocabulary of actions, by simply making additional search engine queries. Experiments show the benefits of this approach in two areas: improving the retrieval precision of human action images, and tagging human actions in YouTube videos. A Multiple Instance Learning framework for exploiting properties of the scene, objects, and humans in video is also proposed for action classification in video.

Third, work towards automatic recognition and retrieval of American Sign Language in video databases will be presented. The effort involves collaboration between computer scientists and linguists. The goal is to enable users to search ASL video content simply by video-recording a query sign and relying on computer-based sign-recognition for lookup. As part of this effort, methods for gesture-based retrieval of signs that can exploit phonological constraints, e.g., on start/end hand shapes in lexical signs, are being developed. The American Sign Language Lexicon Video Dataset (ASLLVD), representing more than 3,300 distinct signs, each produced by 1-6 native ASL signers, for a total of almost 9,800 tokens, is an ASL video corpus that has been gathered and linguistically annotated as part of this effort, and will soon be made available as a resource to the research community.

Collaborators in this work include (in alphabetical order): Vassilis Athitsos, Nazli Ikizler-Cinbis, Hao Jiang, He Kun, Shugao Ma, Carol Neidle, Joan Poole-Nash, Ashwin Thangali, Tai-peng Tian, and others.

Stan Sclaroff founded the Image and Video Computing research group at the University of Boston. He received the PhD degree from MIT in 1995. In 1996, he received an ONR Young Investigator Award and an NSF Faculty Early Career Development Award. Professor Sclaroff has coauthored numerous scholarly publications in the areas of tracking, video-based analysis of human motion and gesture, surveillance, deformable shape matching and recognition, as well as image/video database indexing, retrieval and data mining methods. He has served on the technical program committees of over 60 computer vision conferences and workshops. Stan Sclaroff has served as an Associate Editor for IEEE Transactions on Pattern Analysis, 2000-2004, and 2006-present. He is a Senior Member of the IEEE.

http://dx.doi.org/10.5244/C.26.3