Int J Comput Vis (2012) 100:16–37
DOI 10.1007/s11263-012-0532-9

Coupled Action Recognition and Pose Estimation from Multiple Views

Angela Yao · Juergen Gall · Luc Van Gool

Received: 8 September 2011 / Accepted: 27 April 2012 / Published online: 30 May 2012

Abstract Action recognition and pose estimation are two closely related topics in understanding human body movements; information from one task can be leveraged to assist the other, yet the two are often treated separately. We present here a framework for coupled action recognition and pose estimation by formulating pose estimation as an optimization over a set of action-specific manifolds. The framework allows for integration of a 2D appearance-based action recognition system as a prior for 3D pose estimation and for refinement of the action labels using relational pose features based on the extracted 3D poses. Our experiments show that our pose estimation system is able to estimate body poses with high degrees of freedom using very few particles and can achieve state-of-the-art results on the HumanEva-II benchmark. We also thoroughly investigate the impact of pose estimation and action recognition accuracy on each other on the challenging TUM kitchen dataset. We demonstrate not only the feasibility of using extracted 3D poses for action recognition, but also improved performance in comparison to action recognition using low-level appearance features.

A. Yao (✉) · J. Gall · L. Van Gool
Computer Vision Laboratory, ETH Zurich, Sternwartstrasse 7, 8092 Zurich, Switzerland
e-mail: yaoa@vision.ee.ethz.ch

J. Gall
Max Planck Institute for Intelligent Systems, Spemannstrasse 41, 72076 Tubingen, Germany
e-mail: jgall@tue.mpg.de

L. Van Gool
Department of Electrical Engineering/IBBT, K.U. Leuven, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
e-mail: luc.vangool@esat.kuleuven.be

Keywords Human pose estimation · Human action recognition · Tracking · Stochastic optimization · Hough transform

1 Introduction

Vision-based human motion analysis attempts to understand the movements of the human body using computer vision and machine learning techniques. The movements of the body can be interpreted on a physical level through *pose estimation*, i.e. reconstruction of the 3D articulated motions, or on a higher, semantic level through *action recognition*, i.e. understanding the body's movements over time. While the objectives of the two tasks differ, they share a significant information overlap. For instance, poses from a given action tend to be a constrained subset of all possible configurations within the space of physiologically possible poses. Therefore, many state-of-the-art pose estimation systems use action-specific priors to simplify the pose estimation problem, e.g. (Geiger et al. 2009; Li et al. 2010; Taylor et al. 2010; Lee and Elgammal 2010; Chen et al. 2009). At the same time, pose information can be a very strong indicator of actions and action labels can be determined from as little as a single frame (Schindler and Van Gool 2008; Thurau and Hlavac 2008; Yang et al. 2010; Maji et al. 2011). However, as neither pose estimation nor action recognition are trivial tasks, few systems have tried to couple the two tasks together into a single system. On the one hand, priors from many state-of-the-art pose estimation systems are of a single activity, thereby assuming that the activity is already known, and cannot handle sequences of multiple activities (Taylor et al. 2010). On the other hand, action recognition approaches either model poses implicitly through pose-related descriptors (Thurau