Single Camera Hand Pose Estimation from Bottom-Up and Top-Down Processes

Davide Periquito(✉), Jacinto C. Nascimento, Alexandre Bernardino, and João Sequeira

Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal
davide.periquito@ist.utl.pt

Abstract. In this paper we present a methodology for hand pose estimation from a single image, combining bottom-up and top-down processes. A fast bottom-up algorithm generates, from coarse visual cues, hypotheses about the possible locations and postures of hands in the images. The best ranked hypotheses are then analysed by a precise, but slower, top-down process. The complementary nature of bottom-up and top-down processes in terms of computational speed and precision permits the design of pose estimation algorithms with desirable characteristics, taking into account constraints in the available computational resources. We analyse the trade-off between precision and speed in a series of simulations and qualitatively illustrate the performance of the method with real imagery.

Keywords: Pose estimation · Geometric moments · Hammoude metric · Simulation

1 Introduction

There has been a considerable effort in Human-Computer Interface (HCI) research to create user friendly interaction systems by directly employing the communication and manipulation skills of humans. Adopting such direct sensing in HCI, will permit the deployment of a large spectrum of applications in more complex and sophisticated computing environments such as virtual environments or augmented reality systems. A great focus has been put in the use of hands for interaction devices. The human hand is the most effective interaction tool due to its dexterous functionality in communication and manipulation. A wide range of interaction styles can be reported in the literature. For instance, hand gestures both in static and dynamic settings to build control interfaces [1–3], multimodal user interfaces [4], object manipulation interfaces [5,6] or surgical manipulations [7]. However, to be useful in practice, the interfaces listed

This work was supported by the European Commission project POETICON++ (FP7-ICT- 288382) and the Portuguese FCT projects [PEst-OE/EEI/LA0009/2011] and VISTA (PTDC/ EIA-EIA/105062/2008).

S. Battiato et al. (Eds.): VISIGRAPP 2013, CCIS 458, pp. 212–227, 2014.
DOI: 10.1007/978-3-662-44911-0_14

above need to achieve real-time functionality and precise motion measurement of human hand [8]. Furthermore, they should deal with some challenges among which we highlight: (i) automatic initialisation; (ii) accuracy for long sequences; (iii) independence regarding the activity; (iv) robustness to drift and occlusions; (vi) computational efficiency; (vii) ability to tackle the high dimensionality of the problem (i.e. Degrees of Freedom (DOF)); and (viii) ability to operate with mobile cameras and in uncontrolled environments.

The focus of this paper is the hand pose estimation from a single camera using a pre-trained set of rigid poses. In this context, the hand can be seen as an interaction device with large complexity, with over than 27 degrees-of-freedom (DOF), forming a very effective and general purpose interactive tool for HCI [9]. To achieve this goal we propose a principled way of combining two different sources of information collaborating for the efficient estimation of human hand pose in digital images, herein denoted as *bottom-up* and *top-down*. The *bottom-up* process computes very fast descriptors of the hand pose that, despite their low precision, pre-filter the image information to reduce the computation of the more precise, but slower *top-down* process.

The idea of combining bottom-up and top-down approaches has been successfully exploited in other applications. For instance, in [10], two different methods are used to build models for person detection. First a bottom-up approach searches for body part candidates in the image, which are then clustered to find and identify assemblies of parts that might be people. Simultaneously, a top-down approach is used to find people by projecting the previous assembled parts in the image plane. Other approaches are applied in the context of medical imaging, where the two above mechanisms are combined via online self-retraining [12] and co-training [13] to achieve robustness in the segmentation of the left ventricle from ultrasound images. The combination of bottom-up and top-down processes is crucial for the efficiency and reliability of detection and tracking algorithms. In one hand, the amount of image information to process is huge and thus requires top-down constraints given by models. However, matching the models to the image must be guided by bottom-up processes for efficiency. We evaluate our method with real imagery and study the trade-off between the bottom-up and top-down processes in a series of simulations.

Our paper is organised as follows. Section 2 describes related work. In Sect. 3 we describe the method's architecture, which is divided in to the following major components: (i) the machine learning part (offline) and (ii) the matching strategy between the observed image and the generated hypotheses (online). In Sect. 4 some experiments concerning realistic scenarios are presented. Finally, Sect. 5 presents the conclusions of the paper and provides directions for further research work.

2 Related Work

Two main classes of approaches for hand pose estimation are usually considered depending on the adopted representation for the hand. Considering only part of the hand (i.e. palm, fingers or fingertips) we are facing a *partial* pose estimation. However, if one considers the entire model of the hand, a *full* DOF of the