Multi-view 3D human pose recovery in complex environment
Hofmann, K.M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
1 Introduction

1.1 Motivation

The detection and recovery of humans and their 3D poses in images and video is an important problem in computer vision with many potential applications in diverse fields such as motion capture and analysis for film, interactive games, sports or medical purposes, interfaces for human-computer interaction (HCI), automated surveillance or augmented reality. The ability to recognize humans and their poses is widely recognized as one of the key problems to be solved to enable intelligent man-machine interaction.

In general, the goal of 3D pose recovery systems is to determine the 3D joint locations of a human body from image data. Once recovered, 3D pose can also provide informative, view-invariant features for a subsequent action recognition step. Some applications, for example motion capture-based movement analysis, require detailed pose recovery up to centimeter or millimeter accuracy. To achieve these high levels of accuracy, current commercial motion capture systems typically require the user to wear markers which are then detected by multiple sensors. As the use of these marker-based system is generally quite laborious and costly (application of markers, calibration, etc.), markerless vision-based motion capture systems can provide an attractive non-invasive alternative for many of these applications.

In recent years there has been rapid progress regarding detection and markerless pose recovery of humans in images or video, but successful approaches still mostly rely on accurate object silhouettes obtained using foreground segmentation. Even with ideal scene conditions met in laboratory or studio situations, the pose recovery task is still challenging
due to the articulated structure of human movement. Multiple cameras are placed around the subject – often eight or more – to deal with the underconstrained nature of the problem of recovering three-dimensional information from two-dimensional images due to loss of depth information and/or (self) occlusion. In the last two years, companies offering marker-less human motion capture systems have entered the commercial market. *Organic Motion* was among the first to offer a system which can output, according to the company, “a full 3D model of the subject, complete with surface mesh geometry, surface textures and 3D bone movement data down to millimeter precision.” The promised accuracy and speed can rival that of current marker-based motion capture systems, but the system requires 10 two-megapixel cameras in a controlled-lighting setup, in addition to specialized processing hardware to achieve this performance.

In many practical applications, such ideal scene conditions cannot be met due to the presence of clutter, motion blur and unpredictable background changes, and thus it becomes harder to solve the pose recovery problem robustly. The problem is then often simplified by assumptions on human appearance and motion. For example, subject clothing could be tight-fitting or the articulated human motions might be restricted to particular classes (such as walking, running, etc.). In any case, last years have seen the progression of systems that aim to recover pose on a frame-level (enabling automatic pose initialization) and that combine these estimates with a temporal model to enable tracking.

A recent alternative to the use of video cameras for human pose recovery, is given by 3D cameras, that measure depth directly, either by time-of-flight or by structured light. 3D cameras have proved to be increasingly popular, due to the steady decrease in hardware cost. Their use gained strong momentum since the commercial release of the *Microsoft Kinect* in late 2010, an interactive game console which can track the joints of up to two people simultaneously. Here, the problem is constrained by assuming a front-facing subject at a restricted distance range (few meters). The sensitivity to adverse lighting conditions (bright sunlight) remains to be further investigated.

Over the last years, advances have also been made regarding human shape modeling. What used to be very simple approximations to the human shape (rectangles, cubes, etc.) have become complex parametric volume models or precisely scanned mesh models that approximate
human shape better; this in turn increases the accuracy of model-based 3D pose recovery. However, the increased model complexity comes at the price of increased model acquisition cost, and in many use cases of markerless tracking, acquiring a detailed prior scan of the subject is simply not feasible. Commercially available systems include laser scanners, which can acquire subject data with millimeter precision in about 30 seconds. Generally, the acquired raw data through such systems still has to be transferred to the respective model used for pose recovery. A recent trend has been the joint estimation of pose and shape using a learned, low-dimensional shape model acquired from training data.

Despite all these advances, 3D human pose recovery remains essentially unsolved for unconstrained movement in dynamic and cluttered environments. This holds in particular for scenarios where subjects are outside of the distance ranges required by above mentioned 3D camera systems, or where background and lighting changes complicate recognition. Pursuing this research was guided by the motivation to improve on the state of the art in markerless pose recovery. It is clear that accurate pose recovery necessitates sufficiently accurate assumptions about human shape, and vice versa. In this thesis, we therefore discuss both pose recovery given a suitable shape model as well as shape model adaptation given a generic model and suitable pose initializations.

1.2 Outline and scope

This thesis investigates 3D human pose and shape recovery in “realistic”, complex data, i.e. data that does not exhibit the favorable conditions mentioned in Section 1.1, but potentially includes scene clutter, moving objects such as trains or people, lighting changes due to an outdoor setting, etc. See Figure 1.1 for an example frame depicting a typical setting; note the noisy segmentation and the clutter caused by the presence of moving people in the background. The main focus is on designing a framework for pose recovery, automatic model learning (appearance and shape) and recovery from errors, i.e. enabling automatic pose (re)initialization. To this end, a novel data set was recorded to advance the state of the art in human pose recovery. See Section 2.2.3 for a description. In particular, our chosen setting for the estimation of unconstrained human upper body movement involves the assumption of using only a moder-
CHAPTER 1. INTRODUCTION

Figure 1.1: Example frame (one view shown) with respective foreground mask from foreground segmentation step and pose recovery results overlaid. Note the noisy segmentation and the cluttered background.

ate number of cameras with overlapping field of view, in our case three hardware-synchronized color CCD cameras.

The diversity of human movement encountered in these sequences necessitates fairly weak assumptions regarding the used model of human motion, which in turn attaches greater importance to making likelihood functions and estimation robust. Therefore, we have chosen an approach that detects human pose in every consecutive frame of a sequence and also addresses the above mentioned (re-)initialization problem. Furthermore, the accuracy of our model representation is increased by learning both subject appearance and shape to help in the pose recovery process. In this thesis, an exemplar-based, holistic, representation of human pose was chosen; this has the advantage that upon matching, all available model knowledge or physical constraints can be incorporated at once. A main theme in this work is how to keep the combinatorics of such exemplar-based approach in check, by hierarchical representations or cascaded architectures. Using current PC hardware resources, the exemplar-based approach was scaled up to deal with unconstrained upper-body movement; ways to extend to full-body movement are identified.

A sub-problem of human pose recovery is human detection, i.e. the question of presence of humans in the scene, and at which locations. With our emphasis on pose recovery, we are using a simplified model of detecting subjects in our scene, aided by the use of multiple overlapping cameras and the assumption of only few people in the scene (in our case, not more than two). Generally, the problem of human detection and
1.2. OUTLINE AND SCOPE

tracking rapidly increases in difficulty with an increase in the number of people that are to be detected and tracked reliably; see recent work by Fleuret et al. [2008] or Liem and Gavrila [2009] on the state of the art of tracking small groups (2-6 people).

Due to the complexity of the above mentioned questions, we also decided not to cover topics related to tracking people across non-overlapping cameras (see e.g. work by Zajdel and Kröse [2005] or Metternich et al. [2010]) or related to further semantic interpretation of the poses or movements, such as action recognition or aggression detection [Zajdel et al., 2007]. Human action recognition is an interesting field of its own, but out of scope for this thesis. For more information on action recognition, the reader is referred to surveys on the topic [Aggarwal and Cai, 1999, Gavrila, 1999, Krüger et al., 2008, Moeslund et al., 2006, Poppe, 2010, Turaga et al., 2008, Wang et al., 2003].

The outline of this thesis is as follows. See also Figure 1.2 for an overview diagram of the proposed 3D human pose recovery framework.

Chapter 2 provides an overview over previous work in the field of human pose and shape recovery as well as a summary of our contributions.

Chapter 3 presents a multi-view 3D human pose recovery framework for unconstrained upper-body poses in single frames. The framework is based on a cascaded architecture, where candidate poses are increasingly pruned by successive processing stages. Pose hypotheses are generated and verified by means of probabilistic, hierarchical shape matching.

In Chapter 4, this framework is extended by introducing a temporal integration approach that computes the best trajectories using the provided single-frame detections and a learned motion model. We also use the computed trajectories for the generation of pose predictions which augment the set of detections at the respective next time step.

Chapter 5 deals with two forms of model adaptation. First, we introduce automatic learning of an appearance (texture) model and integrate it into the pose recovery system, for improved performance. Second, we describe an approach to adapt the human shape model based on multiple frames. We address the question which frames should be selected for this shape model adaptation step.

Finally, Chapter 6 concludes the thesis by discussing the strengths and limitations of our approaches and outlines areas of future work.
CHAPTER 1. INTRODUCTION

Figure 1.2: Top-level overview of the proposed 3D human pose recovery framework. See Section 1.2 for further explanation, and Figures 3.1, 4.1, 5.1, 5.9 and 5.20 for refinements of its elements.

The thesis is based on [Hofmann and Gavrila, 2009a,b, 2011a,b].