Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 108 (2007) 4–18
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Vision-based human motion analysis: An overview

Ronald Poppe

Human Media Interaction Group, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands

Received 20 September 2005; accepted 13 October 2006
Available online 25 January 2007
Communicated by Mathias Kolsch

Abstract

Markerless vision-based human motion analysis has the potential to provide an inexpensive, non-obtrusive solution for the estimation of body poses. The significant research effort in this domain has been motivated by the fact that many application areas, including surveillance, Human–Computer Interaction and automatic annotation, will benefit from a robust solution. In this paper, we discuss the characteristics of human motion analysis. We divide the analysis into a modeling and an estimation phase. Modeling is the construction of the likelihood function, estimation is concerned with finding the most likely pose given the likelihood surface. We discuss model-free approaches separately. This taxonomy allows us to highlight trends in the domain and to point out limitations of the current state of the art.

Keywords: Human motion analysis; Pose estimation; Computer vision

1. Introduction

Human body pose estimation, or pose estimation in short, is the process in which the configuration of body parts is estimated from sensor input. When poses are estimated over time, the term human motion analysis is used. Traditionally, motion capture systems require that (electromagnetic) markers are attached to the body. These systems have two major drawbacks: they are obtrusive and expensive. Many applications, especially in surveillance and Human–Computer Interaction (HCI), would benefit from a solution that is markerless. Vision-based motion capture systems attempt to provide such a solution, using cameras as sensors. Over the last two decades, this topic has received much interest, and it continues to be an active research domain. In this overview, we summarize the characteristics of and challenges presented by markerless vision-based human motion analysis. The literature is discussed, with a focus on recent work. However, we do not intend to give complete coverage to all work.

1.1. Scope of this overview

Human motion analysis is a broad concept. In theory, as many details as the human body can exhibit could be estimated. This includes facial movement, movement of the fingers and changes in skin surface as a result of muscle tightening. In this overview, pose estimation is limited to large body parts (trunk, head, limbs). Note that, in human motion analysis, we are only interested in the configurations of the body parts over time and not interpretations of the movement. This means that pose recognition, which is classifying the pose to one of a limited number of classes, and gesture recognition, which is interpreting the movement over time, are not discussed in this overview. For some applications, the positioning of individual body parts is not important. The entire body is tracked as a single object, which is termed human tracking or detection. This is often a preprocessing step for human motion analysis, and we will not discuss the topic in detail in this overview. Surveys of literature on related fields can be found in [78,25] (gesture recognition), and [125] (face recognition).

E-mail address: poppe@ewi.utwente.nl

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.10.016

In the remainder of this section, we summarize past surveys and taxonomies, and describe the taxonomy that is used throughout this overview.

1.2. Surveys and taxonomies

Within the domain of human motion analysis, several surveys have been written, each with a specific focus and taxonomy. Gavrila [27] divides research into 2D and 3D approaches. 2D approaches are further subdivided into approaches with or without the explicit use of shape models. Aggarwal and Cai [4] use a taxonomy with three categories: body structure analysis, tracking and recognition. Body structure analysis is essentially pose estimation and is split up into model-based and model-free, depending upon whether *a priori* information about the object shape is employed. A taxonomy for tracking is divided into single and multiple perspectives. Moeslund and Granum [63,64] use a taxonomy based on subsequent phases in the pose estimation process: initialization, tracking, pose estimation and recognition. Wang et al. [121] use a taxonomy similar to [4]: human detection, human tracking and human behavior understanding. Tracking is subdivided into model-based, region-based, active contour-based and feature-based. Wang and Singh [120] identify two phases in the process of computational analysis of human movement: tracking and motion analysis. Tracking is discussed for hands, head and full bodies.

Currently, we see some new directions of research such as combining top–down and bottom–up models, particle filtering algorithms for tracking, and model-free approaches. We feel that many of these trends cannot be discussed appropriately within the taxonomies mentioned above. We observe that studies can be divided into two main classes: model-based (or generative) and model-free (or discriminative) approaches. Model-based approaches employ an *a priori* human body. The pose estimation process consists of modeling and estimation [100]. Modeling is the construction of the likelihood function, taking into account the camera model, the image descriptors, human body model and matching function, and (physical) constraints. We discuss the modeling process in detail in Section 2. Estimation is concerned with finding the most likely pose given the likelihood surface. The estimation process is discussed in Section 3. Model-free approaches do not assume an *a priori* human body model but implicitly model variations in pose configuration, body shape, camera viewpoint and appearance. Due to their different nature in both modeling and estimation, we discuss them separately in Section 4. We conclude with a discussion of open challenges and promising directions of research.

2. Modeling

The goal of the modeling phase is to construct the function that gives the likelihood of the image, given a set of parameters. These parameters include body configuration parameters, body shape and appearance parameters and camera viewpoint. Some of these parameters are assumed to be known in advance, for example a fixed camera viewpoint, or known body part lengths. Estimating a smaller number of parameters makes the problem more tractable but also poses limitations on the visual input that can be appropriately analyzed. Note that the relation between pose and observation is multivalued, in both directions. Due to the variations between people in shape and appearance, and a different camera viewpoint and environment, the same pose can have many different observations. Also, different poses can result in the same observation. Since the observation is a projection (or combination of projections when multiple cameras are deployed) of the real world, information is lost. When only a single camera is used, depth ambiguities can occur. Also, because the visual resolution of the observations is limited, small changes in pose can go unnoticed.

Model-based approaches use a human body model, which includes the kinematic structure and the body dimensions. In addition, a function that describes how the human body appears in the image domain, given the model's parameters, is used. Human body models are described in Section 2.1.

Instead of using the original visual input, the image is often described in terms of edges, color regions or silhouettes. A matching function between visual input and the generated appearance of the human body model is needed to evaluate how well the model instantiation explains the visual input. Image descriptors and matching functions are described in Section 2.2. Other factors that influence the construction of the likelihood function are the camera parameters (Section 2.3) and environment settings (Section 2.4).

2.1. Human body models

Human body models describe both the kinematic properties of the body (the skeleton), as the shape and appearance (the flesh and skin). We discuss both below.

2.1.1. Kinematic models

Most of the models describe the human body as a kinematic tree, consisting of segments that are linked by joints. Every joint contains a number of degrees of freedom (DOF), indicating in how many directions the joint can move. All DOF in the body model together form the pose representation. These models can be described in either 2D or 3D.

2D models are suitable for motion parallel to the image plane and are sometimes used for gait analysis. Ju et al. [44], Haritaoglu et al. [33] and Howe et al. [38] use a so-called Cardboard model in which the limbs are modeled as planar patches. Each segment has seven parameters that allow it to rotate and scale according to the 3D motion. Navaratnam et al. [70] take a similar approach but model some parameters implicitly. In [40], an extra patch width