Human Motion Tracking and Recognition

Niki Aifanti
Informatics & Telematics Institute, Greece

Angel D. Sappa
Computer Vision Center, Spain

Nikos Grammalidis
Informatics & Telematics Institute, Greece

Sotiris Malassiotis
Informatics & Telematics Institute, Greece

INTRODUCTION

Tracking and recognition of human motion has become an important research area in computer vision. In real-world conditions it constitutes a complicated problem, considering cluttered backgrounds, gross illumination variations, occlusions, self-occlusions, different clothing, and multiple moving objects. These ill-posed problems are usually tackled by simplifying assumptions regarding the scene or by imposing constraints on the motion. Constraints such as that the contrast between the moving people and the background should be high, and that everything in the scene should be static except for the target person, are quite often introduced in order to achieve accurate segmentation. Moreover, the motion of the target person is often confined to simple movements with limited occlusions. In addition, assumptions such as known initial position and posture of the person are usually imposed in tracking processes.

The first step towards human tracking is the segmentation of human figures from the background. This problem is addressed either by exploiting the temporal relation between consecutive frames—that is, by means of background subtraction (Sato & Aggarwal, 2001), optical flow (Okada, Shirai & Miura, 2000), or by modeling the image statistics of human appearance (Wren, Azarbayejani, Darrell & Pentland, 1997). The output of the segmentation, which could be edges, silhouettes, blobs, and so forth, comprises the basis for feature extraction. In tracking, feature correspondence is established in order to locate the subject. Tracking through consecutive frames commonly incorporates prediction of movement, which ensures continuity of motion, especially when some body parts are occluded. For example, when a person is walking, there are some moments when one of the legs occludes the other. So, some techniques try to determine the precise movement of each body part (Sidenbladh, Black & Sigal, 2002), while other techniques focus on tracking the human body as a whole (Okada, Shirai & Miura, 2000). Tracking may be classified as 2D or 3D. 2D tracking consists of following the motion in the image plain either by exploiting low-level image features or by using a 2D human model. 3D tracking aims at obtaining the parameters that describe body motion in three dimensions. The 3D tracking process, which estimates the motion of the body parts, is inherently connected to 3D human pose recovery.

3D pose recovery aims at defining the configuration of the body parts in the 3D space and estimating the orientation of the body with respect to the camera. This work will mainly focus on model-based techniques, since they are usually used for 3D reconstruction. Model-based techniques rely on a mathematical representation of human body structure and motion dynamics. The 3D pose parameters are commonly estimated by iteratively matching a set of image features extracted from the current frame with the projection of the model on the image plane. Thus, 3D pose parameters are determined by means of an energy minimization process.

Instead of obtaining the exact configuration of the human body, human motion recognition consists of identifying the action performed by a moving person. Most of the proposed techniques focus on identifying actions belonging to the same category. For example, the objective could be to recognize several aerobic exercises or tennis strokes or some everyday actions such as sitting down, standing up, walking, running, or skipping.

Next, some of the most recent approaches addressing human motion tracking and 3D pose recovery are presented, while the following subsection introduces some whole-body human motion recognition techniques. Previous surveys of vision-based human motion analysis have been carried out by Cédras and Shah (1995), Aggarwal and Cai (1999), Gavrila (1999), and Moeslund and Granum (2001).

This overview will present some of the techniques proposed in the bibliography, together with their advantages or disadvantages. The outline of this work is as follows. First, a survey about human motion tracking and 3D pose recovery are given. Next, human motion recognition is introduced, followed by a summary of some application works. Finally, a section with future trends and conclusion is introduced.

HUMAN MOTION TRACKING AND 3D POSE RECOVERY

Tracking relies either on monocular or multiple camera image sequences. Using monocular image sequences is quite challenging due to occlusions of body parts and ambiguity in recovering their structure and motion from a single perspective view (different configurations have the same projection). On the other hand, single camera views are more easily obtained and processed than multiple camera views. In Table 1, some recent techniques using only one camera are presented.

In contrast to single-view approaches, multiple camera techniques are able to overcome occlusions and depth ambiguities of the body parts, since useful motion information missing from one view may be recovered from another view. In Table 2, some recent approaches using multiple cameras.

Some currently published papers tackle specifically the pose recovery problem using multiple sensors. A real-time method for 3D posture estimation using trinocular images is introduced in Iwasawa et al. (2000). In each image the human silhouette is extracted and the upper body orientation is detected. Two of the three views are finally selected in order to estimate the 3D coordinates of the representative points and joints. In Rosales, Siddiqui, Alon, and Sclaroff (2001), multiple views are obtained by introducing the concept of "virtual cameras," which is based on the transformation invariance of the Hu moments. One advantage of this approach is that no camera calibration is required.

HUMAN MOTION RECOGNITION

Human motion recognition may be achieved by analyzing the extracted 3D pose parameters. However, because of the extra pre-processing required, recognition of human motion patterns is usually achieved by exploiting low-level features (e.g., silhouettes) obtained during tracking (see Table 3).

APPLICATIONS

Vision-based human body modelling has been used in an extensive spectrum of applications. Some recent works, grouped according to their applications field, are presented in Table 4.

Table 1. Monocular systems

Authors	***Description***
Sminchisescu and Triggs (2001)	A 3D human body model, consisting of tampered superellipsoids, is fitted on the image features by means of an iterative cost function optimization scheme. A multiple-hypothesis approach with the ability of escaping local minima in the cost function is proposed.
Cham and Rehg (1999)	This technique combines a CONDENSATION style sampling with local optimization. A 2D model with underlying 3D kinematics is used. The advantages are its suitability for high-dimensional state-spaces and that the use of discrete features is unnecessary.
Wachter and Nagel (1999)	A 3D model of right-elliptical cones is fitted to consecutive frames by means of an iterated extended Kalman filter. This approach copes with self-occlusions occurring between the legs of a walking person.
Howe, Leventon and Freeman (1999)	Self-occlusions are tackled in this Bayesian system. 3D reconstruction is achieved by establishing correspondence between the training data and the features extracted.
Sidenbladh, Black and Sigal (2002)	A probabilistic approach for modeling 3D human motion for synthesis and tracking, where learning of state transition probabilities is replaced with efficient probabilistic search in a large training set.