Robotics and Autonomous Systems 62 (2014) 818–832

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Motion recognition and recovery from occluded monocular observations

Dongheui Lee [a,*], Yoshihiko Nakamura [b]

[a] Institute of Automatic Control Engineering, Technical University of Munich, Germany
[b] Department of Mechano-Informatics, The University of Tokyo, Japan

HIGHLIGHTS

- 3D whole-body motion recovery from an occluded monocular image sequence.
- Coordinate transformations of statistical database (e.g., HMM parameters).
- A new particle filtering algorithm for estimating baselink position and orientation.
- Concurrent motion recovery and motion recognition.
- Inference from optical flow of feature points.

ARTICLE INFO

Article history:
Received 26 August 2013
Received in revised form
11 February 2014
Accepted 21 February 2014
Available online 3 March 2014

Keywords:
Statistical inference
Motion recognition
Motion recovery
Motion capturing
Optical flow
Particle filter
Monocular vision

ABSTRACT

This paper proposes a method for 3D whole-body motion recovery and motion recognition from a sequence of occluded monocular camera images based on statistical inference using a motion database. In the motion database, each motion primitive (e.g., walk, kick, etc.) is represented in an abstract statistical form. Instead of extracting rich information by expensive computation of image processing, we propose an inference mechanism from low level image features (e.g., optical flow), inspired by psychological research on how humans perceive motion. The proposed inference mechanism recovers the 3D body configuration and finds the closest motion primitive in the motion database. Observations in 2D camera image space can be recognized even though the motion database is prepared in a different space (such as joint space) by coordinate transformation of the statistical motion representation. The approach is view invariant since the demonstrator's baselink position and orientation with respect to camera coordinates are tracked using an extended particle filter. Finally, an experimental evaluation of the presented concepts using a 56-degree-of-freedom articulated human model is discussed.

1. Introduction

Motion understanding of human movements from a camera system which is mounted on a robot is important for realizing smooth and practical human–robot interaction. Although a studio-type motion capture system with several cameras provides good tracking accuracy, the system is expensive and requires a large set up in the environment. Also, human subjects have to wear optical markers on their body and motions can be captured only in the studio. Although a wearable type of a motion capturing system can eliminate space restriction, subjects still have to wear sensors on their bodies. Thereafter, it is inconvenient to use them in daily life environments. Therefore, a new technology for human motion understanding using onboard camera systems seems beneficial for seamless human–robot interaction.

Perception of human motion has been studied in psychology [1–5] in the framework of moving light display (MLD). The moving light display is an experimental setup to show a human motion by lights attached to various parts of the body. These studies report that human can recover and understand three-dimensional human movements from the video while a single static image of the lights is insufficient to find the human shape. The experiments show that humans have high sensitivity to human motion perception and can recover 3D motion from a temporal sequence of images without any structural information. Human motion perception includes spatial and temporal understanding. This suggests that humans use the temporal information and the memory of human motions to recover missing spatial information.

* Corresponding author.
E-mail addresses: dhlee@tum.de (D. Lee), nakamura@ynl.t.u-tokyo.ac.jp (Y. Nakamura).

http://dx.doi.org/10.1016/j.robot.2014.02.002
0921-8890/

With the final goal of capturing three-dimensional human motions and recognizing action classes from an onboard camera system, we focus on an inference mechanism from 2D optical flow without structure information.[1] An approach is proposed to use a human motion database of 3D motion to solve lack of depth information of a monocular camera and occlusion problems. Even a stereo camera system may suffer from depth insensitivity.[2] Although the authors assume a single onboard camera system composed of a monocular camera, the approach can be extended to an onboard stereo vision system. While recently 3D cameras became popular, the proposed technology has benefits for recovering 3D information from 2D video images like film archives as well as smart surveillance systems.

The main contribution of this paper is 3D whole body motion recovery from an occluded monocular image sequence, which includes not only self occlusion but also occlusion by obstacles. The following paragraphs summarize the technical characteristics of the proposed method.

(1) *Coordinate transformation of the statistical database*: In this work, human motion patterns in the database are represented by a time sequence of joint variables and the 3D position/orientation of the basebody[3] to allow for easy control of articulated body motions. For the human motion database, the hidden Markov model (HMM) is adopted [6,7] because it uses a concise representation of spatiotemporal patterns and has well established computational methods. In order to recognize human motions (2D image observations from onboard camera) without the need of a database with many different views, we propose a method to transform the statistical database to an appropriate coordinate. By the coordinate transformation, the HMMs in joint space can be compared with 2D images from any view point without the need of depth information.

(2) *Concurrent motion recovery*[4] *and motion recognition*[5]: One can find many publications of motion recovery [8–10] and motion recognition [11–13] as independent problems. In contrast, our algorithm emphasizes that recovery and motion recognition are tightly coupled in a single framework, where recovery assists action recognition and vice versa. The inference cost for motion recognition in a next time step is significantly reduced by closing the computational loop using recovered motion. Computational concurrency of motion recovery and motion recognition is similar to that of localization and mapping in SLAM (Simultaneous Localization and Mapping) [14,15].

(3) *Inference from optical flow of feature points*: The appearance of people in images varies due to different clothing and lighting conditions [16]. Often used image descriptors include silhouettes [8,17], edges [18,19], color [20], and motion [21,22]. A large computation for image processing of the 2D image sequence would maximally extract information for 3D recognition. Instead, this paper focuses on development of an inference method from low level image features (e.g., optical flow [23] of unlabeled features) without shape and structure information, inspired by human's high perception ability shown in the MLD experiments [1]. Note that the main objective of this paper lies on the inference mechanism from partial monocular observations. In contrast, the reliable feature selection and robust optical flow calculation from blurred images are not the focus of this research. Such methods for image processing (optical flow estimation) can be found in [24,25]. Therefore, to separate these problems in our experiments, we attach artificial markers to the subject as distinctive feature points. Note also that the markers are placed at arbitrary points and neither labeled nor tracked, in contrast to optical markers in conventional motion capturing. Thanks to these properties of random placement of markers, and no need of tracking and labeling, the synthetic observations can be easily replaced with the optical flows from real images. Therefore this allows that the proposed inference method can be directly integrated with 2D optical flows processed from real images.

(4) *Mimesis model*: The basic framework used in this work is the *mimesis model* [6], which was inspired by the mimesis theory [26] and the mirror neurons [27] in cognitive and neuro science. The mimesis model was proposed for imitation learning from human demonstrations, which consists of three components: motion learning, recognition, and generation. This model has been selected because the use of the mimesis model for 3D recovery of human motion patterns may be natural if we recall the fact that our skill of human motion perception is based on tightly connected cognitive activity with learning and reproduction.

The overall data flow is shown in Fig. 1, consisting of the learning procedure and the 3D whole body motion recovery from 2D images. First, during the learning stage, a human performs multiple demonstrations for each motion primitive using conventional motion capturing system. The observed three dimensional Cartesian marker position data $[x, y, z]$ on the human body is converted to joint angle data for a chosen kinematic model[6] using inverse kinematics. The observations in the joint angle space are embodied into the parameters of an HMM (Section 3). The transformation matrix ${}^{C}_{D}T \in SE(3)$ between the camera coordinates and the demonstrator baselink coordinates is found by applying the extended particle filtering algorithm (Section 5.1). In Section 4, the coordinate transformation of the statistical database is described. Motion primitives λ are converted from the demonstrator's joint coordinates λ_θ into the demonstrator's Cartesian coordinates ${}^{D}\lambda_x$ by forward kinematics, into the camera coordinates ${}^{C}\lambda_x$ by a transformation matrix ${}^{C}_{D}T$, and into 2D image Cartesian coordinates ${}^{I}\lambda_x$ by perspective projection. Finally, both proto-symbols ${}^{I}\lambda_x$ and observations ${}^{I}o_x$ are represented in the 2D image Cartesian coordinates. When all the markers are not visible, motion recognition from partial observations are carried out as described in Section 5.2. Section 6 explains how to recover 3D whole body motion close to the 2D observed motion.

Note that there are two stages of motion recovery in this work: one for human baselink position and orientation ${}^{C}_{D}T$ (6DOF) and the other for joint angles (50DOF). The particle filter represents a probabilistic distribution of ${}^{C}_{D}T$ and it influences coordinate transformation and thereafter motion recognition. Motion recognition results affect the prediction of particles at the next step and recovery of joint angles. In this regard, concurrent motion recovery and recognition is implemented in this work.

An earlier version of this work was presented in [28]. This work is extended by in depth explanations of methodology and new experimental results. A method to reproduce a motion sequence by manipulating proto-symbols in different coordinates is newly proposed. While the previous work showed a recovery result of only one occluded motion sequence, this paper provides statistical analysis under different conditions, such as multiple runs with

[1] The kinematic structure of human is invisible.

[2] Even an onboard stereo camera may not achieve complete 3D information of an object far in the distance because of its fixed baseline.

[3] To be precise, our motion database is represented in joint angles, joint velocities, and baselink velocities.

[4] Motion recovery denotes estimation of the sequence of joint angles and basebody position/orientation from the 2D image sequence.

[5] Motion recognition denotes the search for the closest HMM (e.g. walk, run, jump, etc.) to the 2D image sequence.

[6] The kinematic model is chosen depends on an application: for example, a humanoid robot kinematic model for robot imitation of human motions and a human skeleton kinematic model for human motion reconstruction.