ARTICLE IN PRESS

Robotics and Autonomous Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Action database for categorizing and inferring human poses from video sequences

Wataru Takano *, Yoshihiko Nakamura

Mechano-Informatics, University of Tokyo, 7-3-1 Hongo, Bunkyoku, Tokyo 113–8656, Japan

HIGHLIGHTS

- The captured motions are converted into images containing human activities.
- The sequences of images are encoded into the HMMs.
- The images are statistically mapped to the motions.
- The action database contains the motion, image, HMM and mapping.
- The action database enables the action recognition and configuration inference.

ARTICLE INFO

Article history:
Received 31 July 2014
Received in revised form
28 February 2015
Accepted 2 March 2015
Available online xxxx

Keywords:
Action database
Action classification
Hidden Markov model
Gaussian process regression

ABSTRACT

One of the difficulties in automated recognition of human activities is classifying a video into a specific action class by selecting among a large number of human actions. Technology for understanding complex and varied human actions is necessary for automated surveillance, sports training, computer games, and human–robot interactions. The difficulty of classification comes from a dearth of datasets of human actions that are manually categorized and suitable for use as training data for designing action classifiers. A marker-based motion capture system enables precise measurement of human actions for the purpose of analysis. This type of capture system has several drawbacks, however; in particular, marker-based systems are expensive, intrusive, and complex to use. Despite this, the intensive use of a motion capture system can provide large datasets of human actions, and the datasets can be used to facilitate handling the variety of actions to be classified. Large datasets of human actions measured by motion capture systems are expected to be suitable for use in classifying video segments into the correct human action category, selecting from among a large number of action categories, and for inferring human postures from video. This paper proposes a new concept for a database of human whole body actions and an application to understanding human actions from video. The database contains action configurations, such as positions of body parts, pose descriptors from silhouette images, a stochastic model encoding each sequence of the pose descriptors, and a regression model for predicting the configuration from the pose descriptor. The action configurations are recorded in advance of use by measuring many human actions with a marker-based motion capture system, and silhouette images are created from these configurations. We tested the action database on action classification tasks and human body posture inference tasks. The experimental results show that the action database is suitable for use in both action classification and posture inference.

1. Introduction

Action recognition is expected to be integrated into daily life through technologies such as automated surveillance, sign language recognition, computer games, and robot interaction with humans. Action recognition has been the subject of a significant amount of research [1,2]. For the simple case of action recognition, research focuses on classifying videos containing human activity into the correct categories. Action recognition is still difficult when handling a lot of different human activities; however, the aim is not to estimate three-dimensional configurations of humans.

Marker-based optical motion capture systems and RGBD sensor systems are popular ways to measure the positions of human body parts because they are reliable and not affected by cluttered backgrounds or adverse lighting conditions. They have been widely

* Corresponding author. Tel.: +81 3 5841 6378; fax: +81 3 3818 0835.
E-mail addresses: takano@ynl.t.u-tokyo.ac.jp (W. Takano), nakamura@ynl.t.u-tokyo.ac.jp (Y. Nakamura).

http://dx.doi.org/10.1016/j.robot.2015.03.001
0921-8890/

Please cite this article in press as: W. Takano, Y. Nakamura, Action database for categorizing and inferring human poses from video sequences, Robotics and Autonomous Systems (2015), http://dx.doi.org/10.1016/j.robot.2015.03.001

used to measure human actions, and so a large number of action data have been recorded and accumulated [3]. These action data have been used in various applications: for synthesizing character motions [4,5], predicting human actions [6], imitative learning [7,8], and robots' language acquisition from motions [9,10]. However, marker-based optical motion capture systems have several drawbacks. Among these, they are expensive and require users to carefully calibrate the camera system, attach markers to performers, and label those markers. Moreover, the use of marker-based optical motion capture systems is limited to rooms equipped with multiple cameras mounted in fixed positions. To overcome these drawbacks, cheap and useful technology is expected to be developed to recognize human actions and to estimate the configurations of actors for various applications.

We have previously recorded a large dataset of human whole body actions, which we obtained by using a marker-based optical motion capture system, and we have developed a framework for recognizing human actions. This paper describes a novel approach to reusing the action data measured by the capture system, applying it to recognizing human actions and recovering configurations representing the action from a video depicting a human activity. Actions measured by the capture system can be used to create sequences of images recreating that action through rendering a character model performing this action and projecting that model onto the planes of camera images from multiple viewpoints. These sequences of images are encoded into hidden Markov models (HMMs), which are used to classify the action images into the relevant categories. Additionally, the HMMs can be used to group the configurations and images, and an estimator is designed for each group to characterize the configurations from the images. The resulting database of configurations, images, and estimators makes it possible for robots to recognize human actions and recover action configurations from videos of human activities.

2. Related work

Action recognition has been intensively studied and applied to human–robot interaction, computer games, and automated surveillance systems. In the simplest cases, the objective of action recognition is to classify a video depicting human activities into its relevant action category. Each video is a sequence of images, which are sets of two-dimensional data obtained by projecting a three-dimensional scene in the real world onto a plane. For this purpose, video is regarded as a sequence of two-dimensional data, and a sequence of this type is referred to as the space–time volume. A simple method for recognizing actions is to create template images from the space–time volume and then apply a template-matching technique to the video to be categorized [11,12]. The space–time volumes of human actions can be handled as three-dimensional objects. This observation motivates an approach that extracts local features of the space–time volume and recognizes the actions by using the same methods as used for object recognition [13–16]. Several methods have been studied in depth for modeling the spatiotemporal distribution of local features with the aim of improving the performance of action recognition [17–22]. These studies were based on space–time volumes that did not explicitly represent trajectories of joint positions or angles; they focused on action recognition, rather than on recovery of configurations of human actions from video depicting human action. Our interest is not limited to action recognition only.

Trajectory-based approaches are widely used for motion recognition. In these approaches, an action is represented as a space–time trajectory, and more specifically as a trajectory of joint positions or joint angles. These approaches require extracting the joints from each image, but an advantage of trajectory-based methods is that trajectory is insensitive to viewing point. Templates of trajectories [23], or cubic polynomial curves fitted to them [24], have been used to characterize action categories. Gavrila et al. proposed an action recognition method based on dynamic time warping of the trajectories of tracked body parts [25]. The dynamic time warping algorithm searches for the gesture template most similar to the input video, and so it can handle differences in action speed between the template and the input. These trajectory-based approaches require preprocessing to extract and track joint positions before classifying videos into action categories. This preprocessing has a notable effect on the performance of the action recognition.

A variety of parametric approaches, too, have been proposed for action recognition. In these approaches, action categories are represented as parameters of a model trained from the corresponding space–time trajectories. Lublinerman et al. modeled human actions as linear dynamical systems [26]. Sequences of features present in the silhouettes of performers were encoded as the parameters of the linear dynamical system describing the actions. A support vector machine was then used to classify the obtained parameters according to the represented human action. Variant action recognition systems based on linear dynamical systems have been developed for recognizing actions in various situations [27–29]. Probabilistic models are popular for use by parametric representations of actions. In such approaches, a probabilistic model is statistically optimized for each action category in such a way that each action corresponds to the space–time trajectory of its containing category. The probabilistic model for action recognition is most clearly typified by HMMs, which are trained by maximizing the probability that they generate the space–time trajectory of interest. Input videos are then recognized as corresponding to the model that is the most likely to generate the input video. Yamoto et al. proposed a new approach to applying HMMs to recognition of input videos containing actions [30]. In that approach, each image in the video is divided into cells, and the number of pixels of foreground in each cell forms a feature. A sequence of features is then handled as either training data or observation data for the HMMs. HMMs have been extensively used for action recognition systems [31–35]. Parametric approaches such as these are promising frameworks for action recognition. It is worth noting that the action recognition systems described above were tested on a small dataset of actions that were to be classified into one of only a few action categories.

However, integrating an action recognition system into daily life will require a system for learning many action categories and classifying observations into one of the large number of action categories. In our framework, a large number of actions measured by an optical motion capture system create sequences of images containing human activities from multiple viewpoints, and the stochastic relations between the three-dimensional configurations and two-dimensional images make it possible both to recognize actions and to recover the configurations of the actor from only the input images. More generally, this paper contributes to progress on the reuse of databases of space–time volumes and the space–time trajectories for action recognition and human pose recovery.

3. Construction action database

Because marker-based optical motion capture systems have been widely used for various applications, such as video games, character animation, sports engineering, and human–robot interaction, a large number of space–time trajectories (more specifically, sequences of positions of markers attached to a performer) have been recorded and accumulated. The positions of these markers can be converted to sequences of joint positions or angles by inverse kinematic computation using a human figure model. The computed scene of a human figure model performing the specific action is projected onto an image plane, and a sequence of images

Please cite this article in press as: W. Takano, Y. Nakamura, Action database for categorizing and inferring human poses from video sequences, Robotics and Autonomous Systems (2015), http://dx.doi.org/10.1016/j.robot.2015.03.001