Signal Processing: *Image Communication* 42 (2016) 19–30

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Human action recognition using multi-layer codebooks of key poses and atomic motions

Guangming Zhu, Liang Zhang, Peiyi Shen*, Juan Song

[a] *School of Software, Xidian University, Xi'an 710071, PR China*

A R T I C L E I N F O

Article history:
Received 29 August 2015
Received in revised form
11 January 2016
Accepted 11 January 2016
Available online 18 January 2016

Keywords:
Human action recognition
Key pose
Atomic motion
Multi-layer codebook

A B S T R A C T

Taking fully into consideration the fact that one human action can be intuitively considered as a sequence of key poses and atomic motions in a particular order, a human action recognition method using multi-layer codebooks of key poses and atomic motions is proposed in this paper. Inspired by the dynamics models of human joints, normalized relative orientations are computed as features for each limb of human body. In order to extract key poses and atomic motions precisely, feature sequences are segmented into pose feature segments and motion feature segments dynamically, based on the potential differences of feature sequences. Multi-layer codebooks of each human action are constructed with the key poses extracted from pose feature segments and the atomic motions extracted from motion feature segments associated with each two key poses. The multi-layer codebooks represent action patterns of each human action, which can be used to recognize human actions with the proposed pattern-matching method. Three classification methods are employed for action recognition based on the multi-layer codebooks. Two public action datasets, i.e., CAD-60 and MSRC-12 datasets, are used to demonstrate the advantages of the proposed method. The experimental results show that the proposed method can obtain a comparable or better performance compared with the state-of-the-art methods.

1. Introduction

Human action recognition has been a highly active research field of computer vision [3] and robotics [6], including elderly care [8], human–robot interaction [10], etc. With the aging of population, service robots have huge growth potential in home service and hospital service fields [7]. Observing and reacting to human actions automatically is a fundamental skill for service robots [14]. The traditional human action recognition based on the RGB image sequences has been a long-standing challenging work [3,16], partly due to the difficulty to detect and track the human body from RGB images with the absence of depth information. Thanks to the advent of inexpensive depth sensors, such as the Microsoft Kinect sensor, it is possible to capture RGB-D images in real-time. As a result, human action recognition algorithms based on RGB-D image sequences emerge [19]. Taking into consideration the fact that human body is an articulated system of rigid segments which are connected by human joints, human actions can be considered as a continuous evolution of these rigid segments [20]. Therefore, the skeletal data extracted from the RGB-D images can be employed to represent human actions which are less dependent on object affordances [4]. Besides, inspired by [23] in which human trajectories are modeled as heat sources to recognize group activities from the similarity of heat-maps, human actions can be modeled similarly as the trajectories of human joints and the similarity of the trajectories can be utilized to recognize human actions.

* Corresponding author.
E-mail address: pyshen@xidian.edu.cn (P. Shen).

http://dx.doi.org/10.1016/j.image.2016.01.003

However, skeletal data only carry the position and orientation information of human joints in the world coordinate system or in local coordinate systems. Due to the inter-class similarity among different human actions and the intra-class variation when different persons perform the same action, it is not easy to extract robust action patterns for each human action. Furthermore, the stability and accuracy of the skeletal data may be unsatisfactory because of the limitation of the accuracy of sensors and the effectiveness of skeleton extraction algorithms, this will bring negative effect to human action recognition. Under such circumstances, the methods which recognize human actions frame by frame may obtain unsatisfactory results [1,4,8]. Slow feature analysis (SFA) based methods [24] may improve the robustness to intra-class variation, but the inter-class similarity may not be handled evenly. Taking into consideration the fact that one human action can be intuitively represented by a sequence of key poses and atomic motions in a particular order [15], if the key poses and atomic motions of human actions can be extracted simultaneously and separately, the statistical information and the temporal sequence information among key poses and atomic motions may balance the inter-class similarity and the intra-class variation in some degree. Furthermore, the human action representation with key poses and atomic motions can make the traditional periodic motion detection [25] easier in a higher perception level.

In this paper, a human action recognition method based on multi-layer codebooks of key poses and atomic motions will be described. The overview of the proposed method is illustrated in Fig. 1. Inspired by the dynamics models of human joints, we construct feature vectors with normalized relative orientations for each limb, respectively. In order to ensure the effectiveness of key poses and atomic motions, feature sequences are segmented into pose feature segments and motion feature segments, key poses and atomic motions are extracted from pose feature segments and motion feature segments, respectively. Multi-layer codebooks are built based on key poses and atomic motions to represent action patterns of each human action. Besides, left-handed multi-layer codebooks and right-handed multi-layer codebooks are built for each human action both, thus human actions performed no matter by right-handers or left-handers can be recognized straightforwardly without extra mirroring operations. In order to evaluate the effectiveness of the proposal, the Naive Bayes Nearest Neighbor (NBNN) method, the Support Vector Machine (SVM) method and the Random Forest (RF) method are utilized to classify human actions based on the multi-layer codebooks. Two public datasets, i.e., the Cornell CAD-60 dataset [1] and the MSRC-12 dataset [26], are used to verify the performance of the proposed method.

The main contributions of the paper are: (1) We define normalized relative orientation as features for each limb of human body, inspired by the dynamics models of human joints, (2) we segment each feature sequence into pose feature segments and motion feature segments dynamically, based on the potential differences of the feature sequence and (3) we extract key poses and atomic motions from the pose feature segments and motion feature segments respectively, construct multi-layer codebooks with key poses and atomic motions to represent action patterns of each human action, and propose a pattern-matching method to recognize human actions.

The remainder of the paper is organized as follows. In Section 2, the related work about human action recognition are reviewed briefly. Section 3 gives the construction process of the multi-layer codebooks of key poses and atomic motions in detail. The three classification methods are introduced in Section 4. Section 5 presents the experimental results and the discussion about the proposed method. At last, Section 6 gives the conclusion and future works.

2. Related work

2.1. Feature representation

Various kinds of sensor data have been used to recognize human actions, including RGB images [16], RGB-D images [19], depth images [27], skeletal data [28], accelerometer data [29], wireless sensor network data [30], etc. This study focuses on human action recognition methods with skeletal data, so several skeleton-based human action recognition methods will be reviewed briefly in this section. Although the skeletal data may be unreliable due to the occlusion and data corruption, skeleton extraction is not in the scope of this study.

Relative positions or relative orientations of the skeletal joints are usually used for human action recognition. The positions of human joints relative to the torso joint are combined as pose features for each frame and the differences of the pose features in two different frames are calculated as motion features in [1,4,8]. Besides, the rotation matrix of each joint is used to convert the position and orientation in the

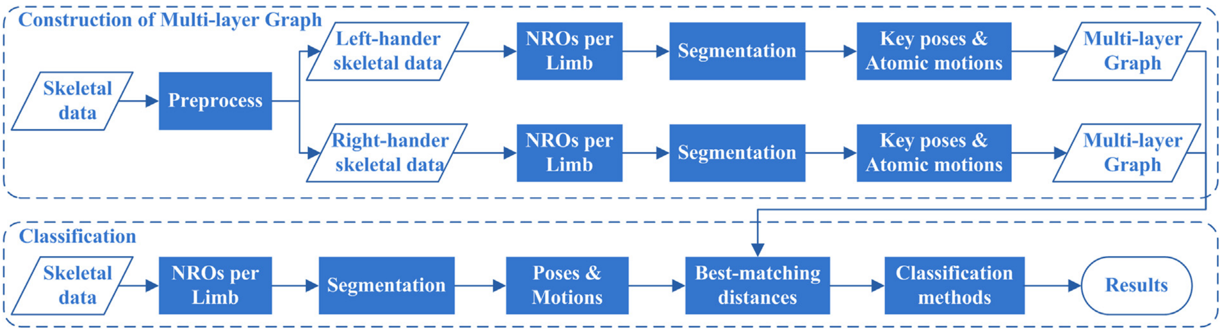

Fig. 1. The overview of the proposed human action recognition method.