Pattern Recognition 61 (2017) 295–308

Contents lists available at ScienceDirect

Pattern Recognition

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

Robust human activity recognition from depth video using spatiotemporal multi-fused features

Ahmad Jalal [a], Yeon-Ho Kim [a], Yong-Joong Kim [a], Shaharyar Kamal [b], Daijin Kim [a,*]

[a] Department of Computer Science and Engineering, POSTECH, San 31, Hyoja-Dong, Nam-Gu, Pohang 790–784, Republic of Korea
[b] Department of Electronics and Radio Engineering, Kyung Hee University, Yongin-si 446–701, Republic of Korea

ARTICLE INFO

Article history:
Received 18 May 2015
Received in revised form
16 May 2016
Accepted 3 August 2016
Available online 4 August 2016

Keywords:
Human activity recognition
Depth silhouette
Skeleton joint extraction
Spatiotemporal multi-fused feature extraction
Hidden Markov model
Forward spotting scheme

ABSTRACT

The recently developed depth imaging technologies have provided new directions for human activity recognition (HAR) without attaching optical markers or any other motion sensors to human body parts. In this paper, we propose novel multi-fused features for online human activity recognition (HAR) system that recognizes human activities from continuous sequences of depth map. The proposed online HAR system segments human depth silhouettes using temporal human motion information as well as it obtains human skeleton joints using spatiotemporal human body information. Then, it extracts the spatiotemporal multi-fused features that concatenate four skeleton joint features and one body shape feature. Skeleton joint features include the torso-based distance feature (DT), the key joint-based distance feature (DK), the spatiotemporal magnitude feature (M) and the spatiotemporal directional angle feature (θ). The body shape feature called HOG-DDS represents the projections of the depth differential silhouettes (DDS) between two consecutive frames onto three orthogonal planes by the histogram of oriented gradients (HOG) format. The size of the proposed spatiotemporal multi-fused feature is reduced by a code vector in the code book which is generated by vector quantization method. Then, it trains the hidden Markov model (HMM) with the code vectors of the multi-fused features and recognizes the segmented human activity by the forward spotting scheme using the trained HMM-based human activity classifiers. The experimental results on three challenging depth video datasets such as IM-DailyDepthActivity, MSRAction3D and MSRDailyActivity3D demonstrate that the proposed online HAR method using the proposed multi-fused features outperforms the state-of-the-art HAR methods in terms of recognition accuracy.

1. Introduction

Recognizing human activities have gained more interest and has become a challenging topic to investigate by the research community. Usually, the analysis is carried out by efficient feature extraction, learning, classification and to compute the input patterns in order to recognize a set of activities. Recently, this research topic has received considerable attention by many real-world applications such as video surveillance system, smart home system, medical health care, sport video analysis and 3D video games [1–4]. In the past decades, a significant amount of research work has been carried out in the field of activity recognition using two categories of sensor devices such as wearable sensors and video sensors. In wearable sensor-based activity recognition, multiple sensors can be attached to a subject's body parts. Most published researches [5–7] described the system based on different sensors as accelerometers, gyroscopes and magnetometers attached on body parts to measure the attributes or features in order to recognize different activities. Unfortunately, it is inconvenient for the subjects to attach the sensors on their body for long time, restricted subject's movements due to wire connections and relatively difficult in terms of energy consumptions and device settings. Moreover, the combination of multiple sensors to enhance the recognition performance may increase the complexity and overall cost of a system, which is inadmissible for real-world applications.

Instead of relying on wearable-based activity recognition, many researches utilized video sensor technologies such as RGB cameras, for activity monitoring and recognition. Chang et al. [8] proposed eigenspace transformation (EST) and Canonical space transformation (CST) methods to extract features from each frame. Then, Fuzzy rule approach is applied on temporal sequence information for recognition and calculate the variation of posture done by different people. Cuntoon and Chellappa [9] introduced an epitomic representation for modeling where the video activity

* Corresponding author.
E-mail addresses: ahmadjalal@postech.ac.kr (A. Jalal), beast@postech.ac.kr (Y.-H. Kim), kimyj@postech.ac.kr (Y.-J. Kim), shka@khu.ac.kr (S. Kamal), dkim@postech.ac.kr (D. Kim).

http://dx.doi.org/10.1016/j.patcog.2016.08.003
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

sequence is divided into segments to extract moving objects and short-time motion trajectories. These information are further processed by Iwasawa matrix decomposition to represent the effect of rotation, scaling and projective action on the state vector and used for activity recognition. Bobick and Davis [10] proposed a view-specific approach for representation of movements as temporal templates. These templates indicate the presence of motion in binary values and the function of the recency of motion in a sequence. Then, a matching algorithm is used to construct a recognition system. Farhadi and Tabrizi [11] extract features as silhouette information, horizontal and vertical optical flow from each frame and stack information from previous and next frames to model local dynamics of activities. Then, the vector quantization used k-mean clustering and 1 Nearest Neighbor (1NN) to recognize an activity. However, these RGB data [8–11] have limited information (i.e., binary or RGB intensity values), difficulty in foreground or background segmentation, motion ambiguities (i.e., color and texture variability) and high sensitivity with lighting conditions which cause major issues to recognize human activities accurately.

With the advancement of imaging technologies and the release of cost-effective depth cameras such as Microsoft Kinect or bumblebee, a new era has started in human activity recognition based on the sequences of depth images. Compared with the digital RGB cameras, depth cameras provide additional body parts information especially during overlapping multiple body parts (i.e., hands in front of chest). In addition, these cameras are insensitive to light changes which enhance performance during darker environments and the body orientation or its size changes are easily normalized which add prominent effects during real-world applications. Basically, several approaches have used depth maps information to explore their features extraction from two basic types such as depth maps-based features and skeleton-based features for recognizing human activities using depth sequences.

In the depth maps-based features, many researches used a set of points of the depth images or human shape silhouettes to extract the features. Oreifej and Liu [12] proposed a new descriptor as HON4D for activity recognition which describes the depth sequences using a histogram capturing the distribution of the surface normal orientation in a 4D space of time, depth and spatial coordinates. Xia and Aggarwal [13] described an algorithm to extract STIPs from depth videos to suppress the noisy measurements. Also, they used a novel depth cuboid similarity feature (DCSF) to describe the local 3D depth cuboid for activity recognition. Jalal et al. [14] developed a novel life logging translation and scaling invariant features approach where 2D maps are computed through radon transform which are further processed to 1D features profile using R transform. These features are further reduced via PCA and symbolized by Linde, Buzo and Gray (LBG) clustering technique to train and recognize different activities based on depth images. Li et al. [15] developed a method based on the expandable graphical model to explicitly model the temporal dynamics of each silhouette and proposed to use a bag of 3D points extracted from the depth maps to recognize the actions. Wang et al. [16] proposed semi-local features called random occupancy pattern (ROP) features, which has employed a novel sampling scheme and extracted from randomly sampled 4D subvolumes with different sizes and locations using depth images. Cheng et al. [17] developed a new depth descriptor as comparative coding descriptor (CCD) which depicts the structural relations of spatiotemporal points within action volume using the distance information in depth data.

Rather depending on depth map-based features, many researches have explored features based on skeleton information. Zanfir et al. [18] developed a moving pose descriptor that considered both pose information as well as differential quantities (i.e., speed and acceleration) of the human body joints. Then, the proposed descriptor is used in conjunction with a modified KNN classifier to consider temporal location of a particular frame for low-latency human action and activity recognition. Sung et al. [19] proposed a set of features such as body pose, hand position, motion information and point-cloud features based on three-dimensional Euclidean coordinates and the orientation matrix of each joint to recognition activities using RGBD images. Yang and Tian [20] proposed a feature based on position differences of 3D joints and Eigenjoints, which include information such as static posture, motion and offset. They used the Naive Bayes Nearest Neighbor (NBNN) for classification. Jalal et al. [21] described labeled body parts approach which requires a dataset of depth silhouettes and their corresponding pre-labeled silhouettes for human body parts generation. The randomly selected features are trained using random forest, whereas the depth intensities pixels are used to generate motion parameters features which are trained and tested using recognizer engine for activity recognition. Xia et al. [22] used a compact representation of postures named HOJ3D that characterizes human postures as histograms of 3D joints locations within a modified specific coordinates system. Then, they trained discrete HMMs to classify sequential postures. Wang et al. [23] proposed spatiotemporal features where 3D joint positions are employed to represent the motion of the human body and local occupancy patterns (LOP) computed interaction between the human subject and the environmental objects based on the 3D point cloud around a particular joint. In addition, Fourier temporal pyramid is used to represent the temporal dynamics in order to improve the accuracy.

In the multi-fused features, some limited researchers have proposed their techniques based on different features combinations (i.e., depth shape plus RGB data or 3D joints plus depth shape). Althloothi et al. [24] developed a method that can characterize both depth silhouettes and 3D joints motion features to represent, classify, and recognize different activities using multi-SVM. Luo et al. [25] proposed a novel framework for depth human action recognition based on RGB and depth features along with Support Vector Machine (SVM). Song and Lin [26] presented three combined RGB-D features, including a local spatial–temporal feature, a skeleton joint feature and a point cloud feature, based on sparse coding to improve the recognition performance. Relying on these existing works proved to be more efficient recognition accuracy using depth sensor. However, it is still difficult to depend on depth-maps features, joint points information or combined features especially during similar postures of different activities and self-occlusions, which badly affect recognition accuracy. Therefore, we needed to develop a novel methodology which provides combined effects of full-body silhouettes and joint points information for human action and activity recognition from depth silhouettes.

In this paper, we proposed new robust spatiotemporal multi-fused features to represent human body shape and recognize human activities using sequence of depth images via the depth sensor. These features are concatenated by two different feature types at the extraction level as skeleton joint features deal with local motion or orientation of joint points information and body shape features deal with intensity difference among human depth silhouettes, respectively. In order to evaluate the performance, a new continuous online human activity dataset is provided that contains segmented video sequences for training and un-segmented video sequences for continuous online activity recognition. It will become a benchmark dataset for continuous activity recognition based on depth data. In addition, we apply the proposed method to public datasets as MSRAction3D and MSRDailyActivity3D datasets and obtain significant improvement of recognition rates over the state-of-the-art methods.

In summary, the major contributions of this paper are