Pattern Recognition 47 (2014) 1800–1812

Contents lists available at ScienceDirect

Pattern Recognition

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

Human activity recognition using multi-features and multiple kernel learning

Salah Althloothi[a], Mohammad H. Mahoor[a,*], Xiao Zhang[a], Richard M. Voyles[b]

[a] *Department of Electrical and Computer Engineering, University of Denver, CO 80208, USA*
[b] *College of Technology, Purdue University, West Lafayette, IN 47907, USA*

ARTICLE INFO

Article history:
Received 17 April 2013
Received in revised form
8 November 2013
Accepted 30 November 2013
Available online 17 December 2013

Keywords:
Human activity recognition
Multiple kernel learning
Spherical harmonics coefficients
Distal limb segments
Support vector machines

ABSTRACT

This paper presents two sets of features, shape representation and kinematic structure, for human activity recognition using a sequence of RGB-D images. The shape features are extracted using the depth information in the frequency domain via spherical harmonics representation. The other features include the motion of the 3D joint positions (i.e. the end points of the distal limb segments) in the human body. Both sets of features are fused using the Multiple Kernel Learning (MKL) technique at the kernel level for human activity recognition. Our experiments on three publicly available datasets demonstrate that the proposed features are robust for human activity recognition and particularly when there are similarities among the actions.

1. Introduction

Human activity recognition has remained as an interesting and challenging topic in the field of computer vision and pattern recognition. This research topic is motivated by many applications such as surveillance systems, video browsing, and human–computer interfaces (HCI) design. In the past two decades, a significant amount of research has been done in the area of human activity recognition using a sequence of 2D images. Most published research is based on either shape features or motion features. Recently, researchers have paid more attention to using 3D spatio-temporal features for describing and recognizing human activities [1–5] due to easy access to depth information via new consumer technologies such as Microsoft's Kinect sensor.

In general, 3D spatio-temporal features look at the changes in the human body shape based on dominant motions in the human limbs [1,3]. The variations in the body shape can be detected and represented with 3D spatio-temporal features as space-time volumes. Those features mainly focus on the representation of the shape and motion as a function of time. The main idea behind the methods that utilize the spatio-temporal features is to recognize human activity by detecting/describing the changes in human limbs either by describing the motion of human limbs or through measuring the similarities among different space-time volumes.

Recently, the developed commodity depth sensors such as Kinect [6] have opened up new possibilities of dealing with 3D data. The Kinect sensor has given the computer vision community the opportunity to acquire RGB images as well as depth maps simultaneously at a good frame rate with a good resolution. As we can see in Fig. 1, the depth map provides additional information as 3D data which is expected to be helpful in distinguishing different poses of silhouettes. Furthermore, compared with RGB images, the depth map increases the amount of information that can be used to detect 3D joint positions.

The research in human activity recognition based on a sequence of depth maps has been motivated with the release of the Kinect Windows SDK, which is utilized to estimate the 3D joint positions of the human body. Although Kinect produces better quality 3D motion than those estimated from regular RGB sensors (e.g., Stereo vision systems for 3D estimation), the estimated 3D joint positions are still noisy and fail when there are occlusions among human limbs such as two limbs crossing each other. Furthermore, the motion of 3D joint positions alone is insufficient to distinguish similar activities such as eating and drinking. Therefore, extra information needs to be included in the feature level to enhance the classification performance. In this context, we need to develop a method to fuse multiple types of features in order to discriminate similar activities and to enhance the recognition rate of the system. For instance, the motion features of human limbs, such as forearms and shins, may be augmented with the shape features that describe the silhouette structure to improve the accuracy of the action classification.

Consequently, fusion techniques can be used to enhance the classification performance of human activity recognition. In this

[*] Corresponding author.
E-mail addresses: salthloo@du.edu (S. Althloothi), mmahoor@du.edu (M.H. Mahoor), xzhang62@du.edu (X. Zhang), rvoyles@purdue.edu (R.M. Voyles).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.11.032

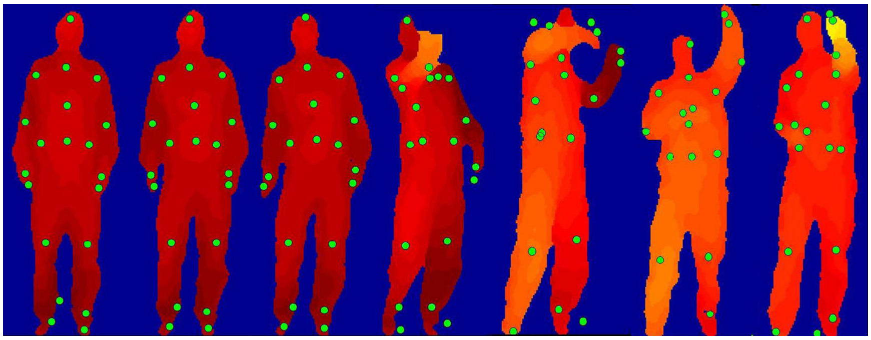

Fig. 1. Sample of depth maps with 3D joint positions for Tennis Serve action.

context, some researchers have conducted fusion at the feature level such as [7–9] where they combined multiple types of features extracted from 2D images into a fused feature vector and then used a single classifier for action recognition. In particular, Liu et al. [7] and Wang et al. [8] combined spatio-temporal volume features and a deformation of the human silhouette obtained from sequences of 2D images to derive action descriptors. In another work, Liu et al. [9] fused local spatio-temporal volumes and statistical models of interest points (Cuboids and 2D SIFT) obtained from 2D images for action recognition using hypersphere multi-class Support Vector Machines (SVM). Fusion can also be performed at the classifier level. For instance [10,11] designed multiple classifiers for two types of features extracted from 2D images, and their final decision was made by taking into account the complementaries among classifiers. In our work, two sets of features are extracted from a depth map (3D data) and are fused at the kernel level instead of the feature level in order to select useful features based on the weights using the MKL technique.

Recently, MKL techniques [12,47] have been proposed for feature fusion within kernel-based classifiers. The works presented in [13–15] show that the MKL technique can enhance the discrimination power and improve the performance of classifiers. The idea behind MKL is to optimally combine different kernel matrices calculated from multiple types of features with multiple kernel functions. Within this framework, the problem of multi-feature representation with a single kernel function in the canonical SVM is transferred to set the optimal value of kernel combination weights for multiple kernel matrices. These works empirically show that the MKL-based multiclass SVM outperforms the canonical multiclass SVM.

This paper presents a method to recognize human activities using a sequence of RGB-D data. The basic idea of our method is illustrated in Fig. 2. Based on the surface representation and the kinematic structure of the human body, we propose a method that can characterize shapes and motions. In our approach the shape features, extracted from the depth map using spherical harmonics representation, are used to describe the 3D silhouette structure. The motion features, extracted from the estimated 3D joint positions, are used to describe the movement of the human body. The distal limb segments of the human body are utilized in our method to describe the motion because we believe that segments such as forearms and shins provide sufficient and compact information for human activity recognition. Therefore, each distal limb segment is described by the orientation and translation distance with respect to the initial frame in order to create motion features. Both sets of features are fused using the MKL technique [16] to produce an optimally combined kernel matrix within SVM for activity classification. This kernel matrix has more discriminating power than a single kernel function due to the utility of multiple features within different kernel functions.

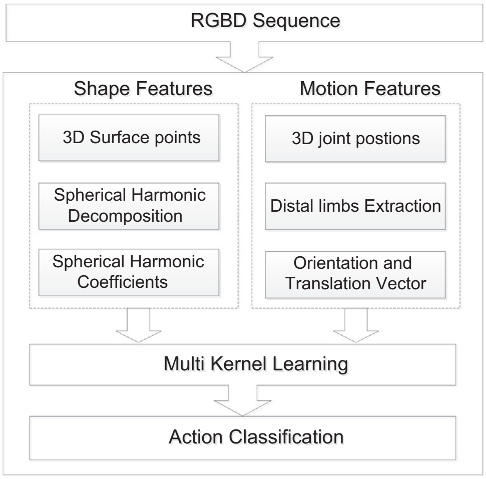

Fig. 2. Our proposed method for human activity recognition using multiple types of features.

Compared with the aforementioned 2D-based approaches for multi-feature fusion, our work is based on features extracted from 3D data (depth map). Also, our approach is based on multiple kernel functions and multiple features which have more advantages over single kernel function with multiple features. In fact, a single kernel cannot perform well when the nature of the features are different and incompatible. Furthermore, combining multiple features into one feature vector introduces the curse of dimensionality problem.

In summary, the contributions of this paper are summarized as follows: (1) A novel 3D shape feature using spherical harmonics transformation to represent the body silhouette is proposed. (2) The human body motions (i.e. kinematic structure) are described using only the distal limb segments. (3) These two types of features are fused at the kernel level as a novel methodology in order to differentiate similar activities and enhance the classification rate of the system.

The remainder of this paper is organized as follows. A brief review of related work is presented in Section 2. Section 3 explains our proposed frame work for activity recognition using 3D spatio-temporal features and feature fusion using the simple MKL approach. Our experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related work

In the last decade, the shape-based methods using 2D images (captured by a regular RGB camera) have been widely used for action recognition. There are several different shape-based