Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 759 – 763

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

3-D Projected PCA based DMM Feature Fusing with SMO-SVM for Human Action Recognition

M. Naveenkumar* and A. Vadivel

National Institute of Technology Trichy, Tamilnadu 620 015, India

Abstract

Action recognition in video sequence is a very important and challenging problem yet. This paper presents an efficient feature extraction method for human action recognition for depth video sequence. For the video sequence acquired by depth sensor, all 3-D projections (xy, yz and zx) are calculated for each depth frame. For each projection view, the difference between each alternative frames have been considered to form the Depth Motion Map (DMM). Principle Component Analysis technique is applied to decrease the facet of DMM-feature. Sequential minimal optimization (SMO) is pre-owned to train the Support Vector Machine (SVM). The proposed approach is evaluated on MSR Action-3D data set and compared with the existing approaches. The empirical results convey that proposed approach achieves good results than the existing approaches.

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Depth Motion Map; Human Action Recognition; SMO-SVM.

1. Introduction

Recognition of human actions in a video sequence with a conventional camera poses a challenging task. Analysis of conventional videos for action recognition has many limitations such as more computation power and also lack of 3D action data. When low cost depth cameras such as Asus Xtion PRO LIVE[1] and Kinect[2] are released in market, research interest on action recognition has been increased. Depth cameras provide the 3D depth data of the object and it will be helpful for segmentation and action recognition. Moreover, depth data is insensitive to illumination changes and also provide the shape information so that it will be helpful for recognition of human actions. The real time applications like smart home and video surveillance applications incorporate the theory of action recognition. There are two main challenges for the human action recognition: 1) description and modeling of the action (Feature Extraction) and 2) Classification of actions.

The contributions in the paper are as follows. Initially in view of computing an efficient technique for feature extraction for human action recognition, fusing sequential minimal optimization learning algorithm is taken for reference. And then the computed results are made suitable for the action inputs taken by depth camera. MSR Action-3D data set is taken for consideration in evaluating the proposed work. The data set comprise 20 action types with 10 subjects. Some of the sample collection from data set is depicted in the Fig. 3. From observations, a positive

*Corresponding author. Tel.: +91 -9791238410.
E-mail address: mnaveenmtech@gmail.com

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.053

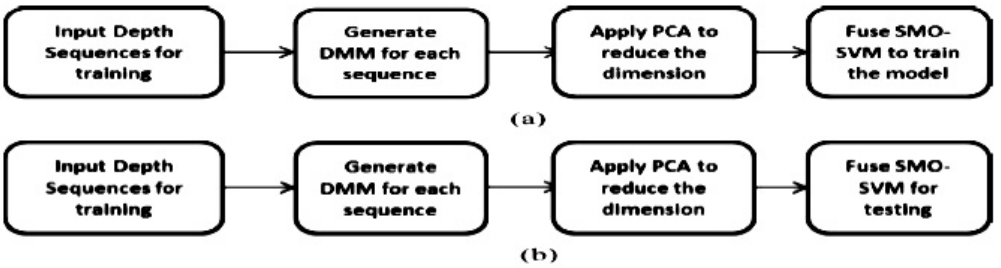

Fig. 1. Work Flows for the Proposed Work: (a) Training Work Flow; (b) Testing Work Flow.

response is seen when compared with the existing ones. The remaining portion in the paper is structured in following manner. The relevant work is reviewed in Section two. The proposed method is delineated in Section three and allied results are reported in Section four. Eventually the paper is culminated in Section five.

2. Related Work

The human action is a spatio-temporal pattern and the representation of feature for action have to grab these patterns. Spatio-Temporal Interest Points (STIP) with Histogram of Gradients (HOG) is used to capture the spatio-temporal pattern[3]. Since depth map doesn't contain texture, this approach fails to classify the actions. Klaser *et al.*[4] explained a spatio-temporal descriptor using 3D-gradients. Actionlet ensemble model is built to depict actions without intra class variance[5]. In literature[6], two features are used: First, a 3D joint feature based on skeleton points captured by depth camera. Second, Local Occupancy Pattern (LOP). After extracting the two features from each frame, temporal dynamics is depicted by fourier temporal pyramid. In literature[7], informative points from the skeleton joints are extracted and used them to form a descriptor for action recognition. Tang *et al.*[8] devised a method to describe a shape of the object efficiently by using normal vectors rather than the gradients. Random occupancy patterns are used for robust action recognition by Wang *et al.*[9]. Oreifej and Liu[10] have introduced a descriptor based on Histogram of 4D normal, HON4D, for perceiving the activities using depth sequences.

An hardware implementation for Histogram of Oriented 4D Normals (HON4D) feature extraction is given by Hsu *et al.*[11] in real time action recognition context. The hardware features less computation and high speed feature extraction with the adoption of sliding histogram concept. The noticeable part of Sliding histogram is that it permits looped classification without video segmentation in advance. In this regard, a bag of 3D points (like a bag of words) from the silhouette of the depth image is recommended by Li *et al.*[12]. Keceli and Can[13] recommended a multimodal approach for action recognition. In this approach, skeleton model extracts angle and displacement features and depth data infers HOG features. To interpret the semantics of actions, the model is trained using Random Forest algorithm. Latterly, Devanne *et al.*[14] recommended a 3-D joint-based framework to seize both the dynamics and shape of the human body simultaneously.

3. Proposed Method

The training and testing frameworks for the proposed work is depicted in Fig. 1. For the video sequence acquired by depth sensor, all 3-D projections (xy, yz and zx) are calculated for each depth frame[15,16]. Unlike literature[15,16], for each projection view, the difference between each alternative frames have been considered to form the Depth Motion Map feature (DMM-f). Assume the depth video sequence contains N frames, the calculation of the DMM-f is as follows. $\text{frame}^i - \text{frame}^{i-2}$ is the depth map of motion energy.

$$\text{DMM-f}_{xy} = \sum_{i=8}^{N-5} (|\text{frame}^i_{xy} - \text{frame}^{i-2}_{xy}|) \tag{1}$$