Expert Systems with Applications 42 (2015) 6957–6965

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Hybrid classifier based human activity recognition using the silhouette and cells

D.K. Vishwakarma *, Rajiv Kapoor

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi 110042, India

ARTICLE INFO

Article history:
Available online 7 May 2015

Keywords:
Human activity recognition (HAR)
Linear Discriminant Analysis
K-Nearest Neighbor
Support Vector Machine
Hybrid classifier

ABSTRACT

The aim of this paper is to present a new approach for human activity recognition in a video sequence by exploiting the key poses of the human silhouettes, and constructing a new classification model. The spatio-temporal shape variations of the human silhouettes are represented by dividing the key poses of the silhouettes into a fixed number of grids and cells, which leads to a noise free depiction. The computation of parameters of grids and cells leads to modeling of feature vectors. This computation of parameters of grids and cells is further arranged in such a manner so as to preserve the time sequence of the silhouettes. To classify, these feature vectors, a hybrid classification model is proposed based upon the comparative study of Linear Discriminant Analysis (LDA), K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) classifier. The proposed hybrid classification model is a combination of SVM and 1-NN model and termed as 'SVM–NN'. The effectiveness of the proposed approach of activity representation and classification model is tested over three public data sets i.e. Weizmann, KTH, and Ballet Movement. The comparative analysis shows that the proposed method is superior in terms of recognition accuracy to similar state-of-the-art methods.

1. Introduction

In recent years, the area of vision based human activity recognition (HAR) has become an important area of research in computer vision, due to its various applications: Surveillance, Assistive health care, Content based video analysis, Interaction between people, Sports, Robotics, and Prevention of terrorist activities (Agrawal & Ryoo, 2011; Chaaraoui, Pérez, & Revuelta, 2012; Vishwakarma & Agrawal, 2012).

The task of the HAR system is to detect and analyse human activity/action in a video sequence. The reviews of previous work (Agrawal & Ryoo, 2011; Poppe, 2010; Vishwakarma & Agrawal, 2012; Weinland, Ronfard, & Boyer, 2011) reveals the challenges in vision based HAR systems. The various factors that make the task challenging are the variations in body postures, the rate of performance, lighting conditions, occlusion, view point and cluttered background. A good HAR system is capable of adapting to these variations and efficiently recognizes the human action class. The important steps involved in HAR systems are usually: (a) Segmentation of foreground (b) Efficient extraction and representation of feature vectors, and (c) Classification. An efficient and novel solution can be proposed at any step of the work individually, or collectively for all the steps. Due to the variation in human body taxonomy and environmental conditions, every step is full of challenges and therefore, one can only provide the best solution in terms of recognition accuracy and processing speed. The Shape and Motion feature based descriptors (Agrawal & Ryoo, 2011) are two widely used methods in HAR systems. Shape based descriptor is generally represented by the silhouette of the human body and silhouettes are the heart of the action. Motion based descriptors are based on the motion of the body, and the region of interest can be extracted using optical flow, and pixel wise oriented difference between the subsequent frames. The motion based descriptors are not efficient, especially when the object in the scene is moving with variable speed.

The main contributions of this paper are twofold: Firstly, for effective representation of human activity, a texture based background subtraction approach is used for the extraction of silhouettes of the human activity from the video sequence. The key poses of the silhouettes are chosen and described by forming cells and grids to produce the descriptor. The modeling of feature vector is done through the easy computation of parameters of grids and cells. This modeling is further arranged in such a manner as to preserve the time sequence of the silhouettes. Secondly, to improve

* Corresponding author. Tel.: +91 11 27871044x1308 (O); mobile: +91 9971339840.
E-mail addresses: dvishwakarma@gmail.com, dkvishwakarma@dce.ac.in (D.K. Vishwakarma), rajivkapoor@dce.ac.in (R. Kapoor).

http://dx.doi.org/10.1016/j.eswa.2015.04.039
0957-4174/

the classification accuracy of HAR system, a hybrid classification model of "SVM–NN" is constructed.

The rest of the paper is organized as follows: Section 2 presents the past work carried out in the field of human activity recognition. The details of the proposed framework which comprises silhouette extraction, feature extraction, feature representation, and various classifiers are presented in Section 3. Section 4 gives the details of experimental work, and the discussion of the result.

2. Related work

Significant amount of work has been reported in the literature for the recognition of human action and activity using video sequences and most of the HAR methods rely on the local features, global features, key points, spatial-temporal features, bags of words etc. (Agrawal & Ryoo, 2011; Chaaraoui et al., 2012; Poppe, 2010; Vishwakarma & Agrawal, 2012; Weinland et al., 2011; Ziaeefar & Bergevin, 2015). All these methods generate a set of features and then an appropriate machine learned classifier is used for the recognition of the activity. A brief review of local feature based spatio-temporal interest point (STIP) and holistic approach that incorporate both local as well as global features are discussed.

An efficient approach of spatio-temporal interest points based on local features using a temporal Gabor filter and a spatial Gaussian filter was introduced by Dollar, Rabaud, Cottrell, and Belongie (2005). Thereafter, a number of STIPs detectors and descriptors have been proposed by several researchers (Chakraborty, Holte, Moeslund, & Gonzàlez, 2012; Everts, Gemert, & Gevers, 2014; Jargalsaikhan, Little, Direkoglu, & O'Connor, 2013; Laptev, 2005; Ryoo & Aggarwal, 2009). These local features based descriptors became popular due to their robustness against noise, illumination change, and background movements. However, these methods seemingly, are less effective for complex activity modeling (e.g. Ballet Movement).

A holistic approach of human action recognition that relies on the human silhouette sequences was proposed by several researchers (Bobick & Davis 2001; Chaaraoui & Revuelta, 2014; Eweiwi, Cheema, Thurau, & Bauckhage, 2011; Gorelick, Blank, Shechtman, Irani, & Basri, 2007; Olivieri, Conde, & Sobrino, 2012; Weinland, Boyer, & Ronfard, 2007; Wu & Shao, 2013). In silhouette based method, the foreground is extracted using background segmentation and then features are extracted from the silhouettes. Bobick and Davis (2001) presented a silhouette based method in which the Motion History Images and Motion Energy Images (MHI, MEI) are used for activity recognition. These MEI and MHI are the images extracted from the video frames and these images are then stacked so as to preserve the temporal content of the activity. Chaaraoui and Revuelta (2014) proposed a method of HAR system that uses the optimized parameters and human silhouette is considered as the basic entity for the feature estimation. The optimized parameters are evaluated using evolutionary computation. Weinland et al. (2007) worked on matching template techniques in which the region of interest (ROI) is divided into a fixed spatial or temporal grid due to which, the effect of noise present in an image and viewpoint variance can be reduced significantly. Thurau and Hlavac (2008) used a histogram of oriented gradients based approach to represent activity, and also concluded that silhouettes are the best information unit for representing human activity. Wu and Shao (2013) proposed modified bags-of-words model called as bag of correlated poses using the advantages of global and local features. They addressed the problem of losing geometric information in bag of visual words representation, which generally is implemented using k-mean clustering algorithm. In these methods, it has been observed that the holistic approach model results in high dimensionality (Agrawal & Ryoo, 2011) of the descriptor; hence there is a need of dimensional reduction techniques for efficient recognition. The PCA is a popular linear dimensionality reduction technique that has been widely used for dimension reduction and classification purpose in activity recognition (Masoud & Papanikolopoulos, 2003; Olivieri et al., 2012). The low dimensional map feature sets are efficiently classified using various linear and nonlinear classifiers. Gorelick et al. (2007) used a nonlinear classification approach for human activity using K-NN along with Euclidian Distance on the global features of the silhouette. Batra, Chen, and Sukthankar (2008) used the nearest neighbor classification approach on local features computed in the form of histograms of code words. Another classification approach that is widely used for human activity recognition based upon local features is SVM and used by Cao, Masoud, Boley, and Papanikolopoulos (2009), Laptev, Caputo, Schuldt, and Lindeberg (2007), and Schuldt, Laptev, and Caputo (2004). Laptev et al. (2007) used the SVM as well as a KNN classifier to classify human activity and showed that SVM gives better accuracy than KNN, but some factors confine their performances like the interclass similarity and intraclass dissimilarity.

Based upon the analysis of earlier state-of-the art methods on human action and activity recognition, we have captured the problems and listed a layout of the solutions of these problems as follows:

- It is observed that the holistic representation of human activity requires an efficient method for the extraction of silhouette from the video sequence. Usually, foreground segmentation is done using background modeling and background subtraction, but it is not always possible to get good results due to inaccurate modeling of the background. Hence, in this work we have used a texture based segmentation approach to extract the silhouette in context of human activity recognition.
- The problem of losing geometrical information in the bags of visual word is addressed by selecting key poses of the human silhouettes. Further, to describe the silhouette information, we have proposed a simple scheme which preserves the spatial change in the human silhouette over time.
- The performance of a classifier reduces when the activities have interclass similarity and intraclass dissimilarity. Therefore, to improve the classification of HAR system, we have constructed a hybrid classification model with the combination of "SVM–NN" classifiers.

3. Proposed framework

Our approach is based on the silhouette of the human body which is extracted from the video sequence of the activity by segmentation techniques. The segmented silhouette is preprocessed to improve its quality for the feature extraction. Features generated from different silhouette images are then arranged in a representable form. Further, dimension reduction, and classification are used. The workflow diagram of the proposed framework is depicted in Fig. 1 and description of each block is presented in following subsections.

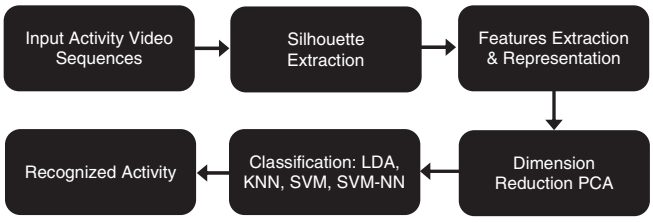

Fig. 1. Workflow diagram of proposed framework.