Int. J. Electron. Commun. (AEÜ) 70 (2016) 341–353

Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

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

Unified framework for human activity recognition: An approach using spatial edge distribution and ℜ-transform

D.K. Vishwakarma*, Rajiv Kapoor, Ashish Dhiman

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

ARTICLE INFO

Article history:
Received 23 March 2015
Accepted 17 December 2015

Keywords:
Human activity recognition
Spatial edge distribution
ℜ-transform
Fusion of translational and rotational features
Still images

ABSTRACT

In this paper, a unified approach for the recognition of human activity using the spatial edge distribution of gradients and orientation of the human silhouettes in a video sequence is presented. The spatial edge distribution is computed on still image at different levels of resolution of sub-images to extract out the shape of the activity posture. The fuzzy trapezoidal membership function is used to extract the key frames of the activity, and the single still key image is extracted according to the histogram distance. The temporal content of the activity is extracted by the computation of orientation of the silhouettes using ℜ-transform. The ℜ-transform is applied on the binary human silhouettes, and the extraction of human silhouettes from the video sequence is done using texture based segmentation techniques. The high dimensionality of the ℜ-transform features is handled by applying Local linear embedding (LLE) dimension reduction approach. A unified model is constructed by integrating the spatial edge distribution of gradients and temporal content of the activity. The performance of the developed model is demonstrated on publicly available datasets, and the highest classification accuracy achieved on each datasets is compared with the similar state-of-the-art techniques and shows the superior performance.

© 2015 Elsevier GmbH. All rights reserved.

1. Introduction

In recent years, human activity recognition has been an active area of research in computer vision due to its potential applications in the field of surveillance, assistive healthcare, sports event analysis, robotics, terrorist activities, content-based video analysis and human-computer interactions [1–3]. However, human activity recognition is both challenging and multifaceted due to viewpoint variations, occlusion, cluttered background, intra-class motion variability and inter-class motion ambiguity. Faced with these challenges several researchers are trying to devise a general, competent and robust method for recognition of human activity.

Over the last few decades, numerous human activity recognition techniques have been proposed, which are mainly focused on the local and global feature based representations. Some of the popular existing approaches for the human activity recognition are as follows; motion flow or optical flow, point trajectories, space-time volume, Bag-of-words model, and spatio-temporal interest points (STIP's) [3–5]. The advantages of these approaches are that they do not require background subtraction and are efficient in handling partial occlusions but all of these techniques have their limitations like optical flow approach is less accurate when the video quality is poor and rough. Similarly, point trajectories based approach requires an efficient tracking of the human motion and if human is moving with variable speed then tracking trajectories may be inefficient. In case of STIPs the distribution of the interest points should be stable around the object. Even the bag of words approach is inadequate to capture the spatial and temporal information and only focus on the global saliency and ignores the structure of the body.

In these days, a trend in the human activity recognition (HAR) has been realized, where multiple features [6–8] are used to improve the recognition accuracy. These methods include global and local information and admit that an individual feature based methods are less effective as compared to the multiple features based methods. In this context more recently, researchers [9,10] supported that multiple features based fusion techniques can provide better performance than the individual features. Wu and Shao [12] presented the combination of local and holistic representation for human action recognition where they effectively worked on the Bag of correlated poses for the local representation and involved MHI/GEI images for the holistic representation. This combined approach is robust to viewpoint, scale and orientation but has plenty of scope for improvement using multiple features. Liu et al.

* Corresponding author. Tel.: +91 1127871044x1308/9971339840.
E-mail addresses: dvishwakarma@gmail.com, dkvishwakarma@dce.ac.in (D.K. Vishwakarma), rajivkapoor@dce.ac.in (R. Kapoor), ashish.dhiman1@gmail.com (A. Dhiman).

http://dx.doi.org/10.1016/j.aeue.2015.12.016
1434-8411/© 2015 Elsevier GmbH. All rights reserved.

[13] proposed the adaptive learning methodology where they used the genetic programming for the representation of spatio-temporal features. Simultaneously, they fused the color and motion information for high-level activity recognition. The main drawback was that they required a large number of generations for good results that further lead to heavy computation. Liu et al. [14] give a probabilistic model, which effectively combined the Gaussian process (GP) regression with sparse covariance matrix for realistic action recognition. However, GP is limited to large-scale computer vision tasks because modeling the large dataset with stochastic process remain a challenge. Hu et al. [15] introduced the spatial pose based exemplars to characterize the Human–Object Interaction (HOI) from still images but it did not work accurately for complex images. This approach does not provide the motion information in the short or long duration of time and hence, scarce in representing the human action from a video sequence. Shao et al. [16] presented the content based search algorithm for localization of human action in the video database. Their work mainly focused on the temporal and spatial localization to decrease the search time of the algorithm. Even though it has a limitation with the large online database, it opens the window for a robust and effective content-based searching algorithm with the existing human action recognition approaches.

Recently, the concept of still images [17–23] has emerged as a popular means for detecting a person's activity or behavior. In these approaches the visual appearance of the object is used to describe the information content. Wang et al. [19] introduce the concept of action recognition based on still images. The shape of human action is represented using a Canny edge detector, and similar body poses are clustered using the spectral clustering method. Li and Ma [20] gave "exemplarlet" based feature descriptor that contains enough visual information to identify in still images. Li and Fei-Fei [21] presented an integrated method that is based on the appearance information on still image and occurrence of action scenes. Thurau and Hlavac [24] proposed the method based on pose information of human action, where they determined the region of interest (ROI) images and further calculated the histogram of gradients with non-matrix factorization to represent the feature vectors. Lopes and Santos [22] proposed the transfer learning approach, where contextual information is extracted from the still images, and a similar approach is used by Zheng et al. [23] for the representation of human action by combining the poselet with contextual information. But in general it is observed that still image based human activity recognition is less effective due to missing temporal information. Hence, more holistic solution is that which utilizes the shape as well motion information together.

Motion temporal information at different orientation is extracted by using the $\Re$-transform [25–27] and extensively used for representing human activity. The $\Re$-transform is applied on the silhouettes of the human body and provides the orientation of the silhouettes. Also, the rate of change of orientation of the human body is different for dissimilar activity. Wang et al. [19] use $\Re$-transform to represent the low-level features due to its advantages of low computational complexity, geometrical invariance. Zhang et al. [25] used a simple approach to the representation of the human activity by using the shape information. They used the $\Re$-transform as a shape descriptor and reported that it works better for the activities that are being performed by the rotation of the human body. Similar work based on $\Re$-transform by Khan et al. [26] for representation of abnormal human activities were carried and it was observed that it is a good descriptor for the orientation based human activity. The properties of $\Re$-transform i.e. invariant to translation, scaling [28] and effectively depicts the change in rotation makes robust feature descriptor.

Nevertheless, most of the reviewed work reveals that an individual feature descriptor has advantages of their own as a single still image based feature representation does not require any background subtraction, morphological operation, or tracking trajectories, thus reducing the computation time and complexity of the system. Therefore, it can be said that these images become, occlusion free and robust to noise. But apart from these benefits, it is also observed that a single still image based technique requires effective positioning of the posture, and it alone does not always provide enough information for recognizing all kinds of activities because it does not contain the temporal information in short or long time duration. In earlier works, it has been observed that $\Re$-transform is effectively used to incorporate the spatio-temporal content of the human activity and found more effective for the abnormal activities representation as compared to the normal activities. In general, human activities are performed by the translation and rotation of human body, and the normal activities have more translation than the rotation while abnormal activities have more rotation than translation.

More recently, several researchers [6,7,10,11] have promoted that multiple feature based fusion techniques give better performance than the individual feature based technique. Hence, in this work a unified structure is proposed by considering the facts of action dynamics. The action dynamics of the human body state that an activity cannot be accomplished without translation and rotation characteristics of the human body. Therefore, for effective representation of human action, an effective descriptor may be formed by incorporating these two characteristics, which could be occlusion free, robust to noise, and computationally less complex.

In this paper, multiple features based integrated structure is proposed where the shape of human pose is recorded in single 2D posture and the motion of the human body is recorded in human silhouettes. The translation provides the change in shape, and sequence of orientation provides the nature of the activity. Due to the translation, the appearance of 2D human postures of different activities is different, and the single key pose is chosen, which have highest variations among the postures of the video sequences. The rotation of key binary human silhouette is computed by applying $\Re$-transform. The binary human silhouette is obtained using texture based foreground segmentation, which is a robust and reliable approach to the illumination change and noise [29]. The key contributions of the work are as follows:

- The appearance of 2D human pose is chosen from the video sequences using fuzzy logic based model. The edge of human body pose is extracted using canny edge detector.
- The edge spatial distribution of gradients at various orientation bins is computed at different sublevels of the single 2D posture for the representation shape of activity.
- The spatio-temporal motion content of human activity is extracted by applying the $\Re$-transform on the key poses of the binary human silhouette. The key poses of binary human silhouettes are chosen on the basis of high energy.
- An integrated model is constructed using the 2D shape information and spatio-temporal motion information of the human activity.
- The performance of the integrated model is measured on publicly available standard datasets using K-nearest neighbor (K-NN) and support vector machine (SVM).
- The comparative analysis of the result achieved through the proposed model is done with the earlier state-of-the-art methods, and an additional analysis of the speed of computation and robustness test of the proposed algorithm is done.

The rest of this paper is structured as follows: Section 2 gives the proposed methodology, which includes the overview of proposed model, extraction key pose, fuzzy logic model, abstraction of spatial edge distributions and motion temporal information using $\Re$-transform, and silhouette extraction. Section 3 gives the details

Download English Version:

https://daneshyari.com/en/article/446106

Download Persian Version:

https://daneshyari.com/article/446106

Daneshyari.com