Optik 126 (2015) 3737–3742

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

View invariant real-time gesture recognition

Somsukla Maiti [a,b], Sandeep Reddy [c], Jagdish Lal Raheja [a,b,*]

[a] Academy of Scientific and Innovative Research, New Delhi, India
[b] CSIR-Central Electronics Engineering Research Institute, Pilani, Rajasthan, India
[c] Birla Institute of Technology, India

ARTICLE INFO

Article history:
Received 16 September 2014
Accepted 30 August 2015

Keywords:
View-invariant
Gesture recognition
Hidden Markov Model (HMM)
Kinect
Angular orientation

ABSTRACT

The invention of several machine assisted systems has increased the scope of security in the society. Human activity recognition is one such very useful human computer interactive system that has gained the attention of the researchers for decades. By identifying and recognizing different actions performed by people, proper surveillance of different criminal activities or activities of elder people staying in the home alone can be performed. This paper as a whole contributes to this cause by developing a robust technique to identify the activities performed by person in a room. This technique provides a solution to most of the previous camera conscious approaches, by generating a system prototype that is view-invariant. For this purpose multiple Kinect sensors have been used. The proposed system is fast and accurate under different illumination conditions and the activity is recognized instead of the fact whether the person is looking at any camera or not.

1. Introduction

Human activity recognition is an important area of research in the field of artificial intelligence. The activities by a human being can tell a lot about the physical condition of the person. So it can be used explicitly in different surveillance applications like the military surveillances, health monitoring, etc. This can be used in the field of healthcare and patient care systems [1,2]. The major concern of the patient monitoring systems, and eldercare systems are to monitor the physical condition of the aged and sick people and to take proper steps to prevent accidents. Most of these applications require an automated recognition of the high-level activities, composed of multiple simple (or atomic) actions of persons.

Activity recognition basically refers to the recognition of sequence of actions performed over a certain period of time. The representation of a single activity using a set of identifiers is a major factor for the activity recognition. Different cues, used for activity recognition, include sensor data obtained from the wearable sensor methods [3–5] as well as soft biometric traits and other features obtained through the view based approaches [6,7]. The recent and most used approach of representation of any human action is performed by considering the depth view of the person. The launch of depth cameras like Microsoft Kinect, ASUS Xtion PRO has made these works quite easy. Microsoft Kinect was introduced in 2010 which is cost effective and provides depth maps at a rate of 30 frames per second (fps) along with the RGB data. Because of the low response time and good performance of the camera, both with the depth data and the colour data, the use of this camera is preferable. Thus most of the researchers have an inclination towards using Kinect [8,9] skeletal representation for the gestures.

Most of the systems, till day, rely on single-view video information of the gestures. The single-view approaches often suffer from ambiguities and self-occlusion. It forces the users to perform the actions facing the camera only. In the real-life situations, it is not possible to face the camera, every time while performing any action. Specifically in situations, where the aim is to monitor the activities of people in a room, the possibility of performing the actions facing the camera will be very small. To make the system robust, the idea is to design a system that is view-invariant of the camera [10]. To increase the reliability of the view invariance of the system, the proposed system employs more numbers sensors for this task. In this paper, we have tried to recognize full body gestures like standing, kicking, bending, etc., using the depth videos generated using multiple number of Microsoft Kinect sensors, by providing a view-invariant setup.

2. Literature survey

Human activity recognition is quite a popular research topic in the area of human computer interaction. Earlier researches on

* Corresponding author. Tel.: +91 1596252444; fax: +91 1596242393.
E-mail address: jagdish.raheja.ceeri@gmail.com (J.L. Raheja).

http://dx.doi.org/10.1016/j.ijleo.2015.08.243
0030-4026/

activity recognition showed notable successes. Different end applications, like health monitoring [11], navigation system for people with cognitive disabilities systems [12], etc., have been developed recently and several further end applications can be developed in this field. Different machine learning techniques including decision trees, Bayes classifiers, and nearest-neighbor algorithms have also been explored for activity recognition [13,14].

In the earlier researches, most of the activity recognition was based on the 2D videos [15]. One of the most common approaches in this field is the consideration of the space-time feature model corresponding to the point of interest in the sequence of frames in video. Other methods of recognition, including the template matching approaches, state-space approaches, semantic description of human behaviours, that have been observed earlier for gesture recognition purpose [16]. These methods work only fine for the 2D videos or still images. But the problem occurs if there lateral changes in the action sequences.

The method of using depth images and further skeleton extraction [9] has been adopted by Jiang et al. in 2010. Consequently, the 3D joints [17,18], including the depth coordinates, of the skeleton are used. Several template based approaches [19,20] provides good result for gesture recognition, but due to the large storage of all the templates, the template based methods are not suitable and not recommended.

Activity recognition using the Kinect has been tried earlier [21] and the result shown was satisfactory to move the work forward in the same direction with further modifications. To perform gesture recognition with 3D joint data, eigen joint approach [22] was followed and it provides quite good result in recognition. But the eigen joint considers to many joints and thus the method is computationally heavy. Bag of features approach has been used earlier for the 3D joints [23]. The performance of the system is very much user dependent and the accuracy changes a lot if any activity is cross-checked by another user.

The hierarchical action sequence of any gesture can be modelled using hierarchical probabilistic graphical model. A hierarchical method for activity recognition has been performed using the process of two layer hierarchical Maximum Entropy Markov Model (MEMM) [24]. This model assumes that any activity occurs in a unique sequence and is repeated every time. The performance of the algorithm is obtained as 84.3% in case where the person was there in the training database because of the consideration of only RGB data of the person. The system does not work properly if the object gets occluded.

Another very important technique that has been used extensively for activity recognition involves the concept of Hidden Markov Model (HMM) [25]. The concept of HMM has been adopted to recognize two-handed activities [26]. The performance in [24] has been improved a lot by using a method that takes the histogram of the 3D joints of the skeleton obtained from the depth data of the person has been followed in [18]. Different feature extraction approaches have been adopted in order to improve the rate of recognition. It also includes the features extracted out of the 3D joint data [27]. The introduction of the HDP–HMM models [28] and further classification by means of one-class SVM provided a better option for the classification of the activities.

A method called Histogram of Oriented 4D Normals was discussed in [29]. This approach included the 3D locations of the joints as well as the motion of the joints. But as the motion of any joint can be similar for different activities, the method is not reliable.

Li et al. has proposed a method [30] to employ an action graph that has been obtained from the video sequence of dynamic movement of body parts. A bag of 3D points by sampling points from the silhouette of the depth images has been obtained and, then clustering the points in order to obtain salient postures (vocabulary) is performed. Consequently, a Gaussian Mixture Model (GMM) [31,32] is used to globally model the postures, and the action graph is used for inference. But this becomes quite more complex for simple activities and is quite time consuming as it is following GMM.

Quite interesting approach of activity recognition was developed based on the action trait code [33]. The average velocities of different body parts were taken as small elements of action and they are termed as action trait elements. The codebook generation was based on the features called the elements and simple classification approaches were adopted to classify the actions. While using a random large set of actions, the cokebook generated becomes quite big and thus the approach becomes computationally heavy. Similar kind of work has been performed that consider each signification elementary action as keyframes [34] and they were encoded using a spatially localizable poselet-like representation with HoG and BoW components learned from weak annotations. The learning and classification is solely based on structured SVM [35,36] and thus it is too much parameter dependent.

A view-invariant multimodal [37] tracking system has been proposed that includes accelerometers, stereo vision camera 2 Kinect, mikes, etc. The system considers the variance of joint angles to form the feature set and perform the classification using SVM. But inclusion of separate mikes and cameras makes the system more costly.

A very useful method was being suggested in [38], which include all the simple sequence of actions to perform any activity. Thus an action bank is formed and template-based action detection is performed. The method is invariant to changes in appearance, but the storage of the action sequence is quite high. Similar approach of collecting the features for each gesture and then generation of the codebook [39] has been adopted in 2003. The approach too is totally view-dependent and demands full attention of the user.

3. Proposed method

The activity recognition system is mainly aimed to monitor the activities of any person in a room or any bounded area. The multiple-view provided by multiple Kinect [40] sensors, placed at different angles to each other, provides a better opportunity to solve the problem as described earlier. The use of the multiple cameras provides the advantage of performing the actions facing any direction in the room. The idea is to install multiple cameras facing different directions in the room and then to decide upon the camera towards which the person is currently facing or the camera with which the person makes the least angular rotation. Once the camera gets selected, the further process of training and testing is performed using the selected camera only. This makes the system capable of recognising the activity, independent of the cautiousness of the person. On the other hand, the system will also be capable to focus on the context of the situation of the activity.

3.1. Experimental setup

The arrangement of the Kinect sensors for the robust setup is given in Fig. 1. Considering the view angle and the IR sensing limitations of the Kinect and the dimensions of the room, we can decide upon the number of Kinects to be placed in the room and the proper location of the Kinects. The viewing angle of a Kinect is given as 43° vertical by 57° horizontal field of view. Thus the number of Kinects to be used is calculated as the coverage of angular view of the Kinect.

3.2. Selection of sensor/sensors

The idea of the proposed system is to find the angular position of the skeleton with respect to the straight line view of the camera. The cameras are placed at certain angles to each other as shown