UNUSUAL EVENT DETECTION IN VIDEO SURVEILLANCE TO PREVENT NUISANCE IN SOCIETY USING COMPUTER VISION APPROACH AND MACHINE LEARNING

[1]Purav Manoj Shah, [2]Sai Kailash

[1,2] Electronics and Communication Engineering (ECE) Department, B.M.S. College of Engineering, Bengaluru, Karnataka - 560019, India

Abstract - Humans can easily understand actions in a complex scene by using a visual aid. One of the main aims of this research paper is to make machines analyze and recognize human actions using motion information as well as different types of information. One of the challenging issues is the process of recognizing and understanding human actions from videos owing to large variations in human appearance, pose changes, and scale changes. The most important approach for human action recognition is to extract features from videos as representations. It is a main area of the computer vision approach. The main applications include surveillance systems, patient monitoring systems, and several systems that involve interactions between persons and electronic devices such as human-computer interfaces. Almost all applications require an automatic recognition of high-level activities.

Keywords *- Fuzzy Inference System (FIS), Kinect Camera, Action Recognition, Surveillance*

I. INTRODUCTION

Humans can easily understand actions in a complex scene by using a visual aid. One of the main aims of this research is to make machines analyze and recognize human actions using motion information as well as different types of information. There are four important processing stages present in an action recognition system; they are human object segmentation, feature extraction and representation, activity detection, and classification algorithms. Three main action recognition systems are present; single person action recognition, multiple person action recognition, and abnormal action recognition and crowd behavior. There are four stages for action detection. The human object is segmented out from the video series first [1]. The different features of the human object such as shape, silhouette, colors, poses, and body motions are then properly extracted into a set of features. Thereafter, an action detection or classification algorithm is applied to the features that are extracted to recognize the various human activities. The starting and ending times of all occurring activities from an input video must be detected for the recognition of human actions [1]. Several important applications can be constructed for the recognition of complex activities. Automated surveillance systems in public places such as airports, railway stations, bus stations, and stadiums. require the detection of abnormal and suspicious activities as against normal activities. For example, an airport surveillance system must be able to automatically recognize suspicious activities like “an individual leaving his/her luggage” or “an individual placing his/her luggage in a dust bin” [1].

Recognition of human actions makes it possible to monitor patients, children, and elderly persons in real-time. There are various types of human actions present. Depending on their complexity, they generally categorize human actions into four different levels: gestures, actions, interactions, and group activities [1]. These complexities make the research topic application-oriented and challenging.

There are different action recognition systems present and some are reviewed here. Action recognition using sparse representation is a recent method, these representations can be built by decomposing signals over elementary waveforms which are chosen from a family called a dictionary. Signals carry a large amount of data that contain both relevant and irrelevant information where relevant information is difficult to be obtained. In sparse representations, few coefficients contain the relevant information. These representations can improve pattern recognition, compression, and noise reduction. Also, they are robust against missing data and distortions which finally provides a compact representation useful for action recognition.

II. LITERATURE SURVEY

Over the past decade, a great deal of work has been done on recognizing human activities. However, the problem is still open and provides a significant challenge to the researchers, and more rigorous research is needed to come around it. An overview of the various action recognition methods and available well-known action datasets are provided in Taha et al. [2]. Previous research in gesture recognition was based on color or grayscale intensity images. These images are obtained from traditional RGB cameras, where each pixel's value represents the intensity of incoming light. It contains rich texture and color information, which is very useful for image processing. However, it is susceptible to illumination changes.

Recently, vision technologies can capture distance information from the real world, which cannot be obtained directly from an intensity image. These images are obtained from depth cameras, where the value of each pixel represents the calibrated distance between camera and scene. An advantage of using these sensors is that they give depth at every pixel, so the object's shape can be measured. While using depth images, computer vision tasks like background subtraction and contour detection become easier. There are many signs of progress, and improvements have been made with the use of depth information.

Based on the paragraphs above, there are two main approaches for human behavior recognition: RGB video-based approach in [2] and depth map-based approach in Ye et al. [3], Chen et al. [4]. This section focuses only on reviewing the state-of-the-art techniques that investigate the applicability and benefit of depth sensors for action recognition, especially skeleton-based approaches. The use of the different data provided by the RGB-D devices for human action recognition goes from employing only the depth data, or only the skeleton data extracted from the depth data, to the fusion of both the depth and the skeleton data. Existing skeleton-based human action recognition approaches (Vemulapalli et al. [5]) can be broadly grouped into two main categories: joint-based approaches and body part-based approaches. Joint-based approaches consider the human skeleton as a set of points, whereas body part-based approaches consider the human skeleton a connected set of rigid segments. Approaches that use joint angles can be classified as body part-based approaches since joint angles measure the geometry between directly connected pairs of body parts.

Jalal et al. [6] present a depth-based lifelogging human activity recognition system to recognize the daily activities of older adults and turn these environments into an intelligent living space. Initially, a depth imaging sensor is used to capture depth silhouettes. Based on these silhouettes, human skeletons with joint information are produced, which are further used for activity recognition and generating their life logs. The lifelogging system is divided into two processes. Firstly, the training system includes data collection using a depth camera, feature extraction, and training for each activity via Hidden Markov Models. Secondly, after training, the recognition engine starts to recognize the learning activities and produces life logs.

Gasparrini et al. [7] propose an automatic fall detection method using the Kinect depth sensor in the top-view configuration. Their approach allows detecting a fall event without relying on wearable sensors and by exploiting privacy-preserving depth data only. Starting from suitably preprocessed depth information, the system can recognize and separate the still objects from the human subjects within the scene using an ad-hoc discrimination algorithm. Several human subjects may be monitored through a solution that allows simultaneous tracking. Once a person is detected, he is followed by a tracking algorithm between different frames. The use of a reference depth frame, containing the setup of the scene, allows one to extract a human subject, even when he/she is interacting with other objects, such as chairs or desks.

Althloothia et al. [8] present two sets of features for human activity recognition using a sequence of RGB-D images: shape representation and kinematic structure. 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 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.

Wang et al. [9] present an Actionlet Ensemble Model for human action recognition with depth cameras. An action let is a particular conjunction of the features for a subset of the joints, indicating the features' structure. As there is an enormous number of possible actionlets, the authors propose a data mining solution to discover discriminative actionlets. An action is then represented as an Actionlet Ensemble, which is a linear combination of the actionlets, and their discriminative weights are learned via a multiple kernel learning method.

Ofli et al. [10] propose a skeletal motion feature representation of human actions, called Sequence of the Most Informative Joints (SMIJ). Specifically, in the SMIJ representation, a given action sequence is divided into several temporal segments. Within each segment, the joints that are deemed to be the most informative are selected. The sequence of such joints is then used to represent an action. One of the limitations of the SMIJ representation that remains to be addressed is its insensitivity to discriminate different planar motions around the same joint. The joint angles are computed between two connected body segments in 3D spherical coordinates, thus capturing only a coarse representation of the body configuration.

III. PROPOSED SOLUTION

Video surveillance has attracted a lot of attention from the computer vision community in recent years. The increasing demand for safety and security has resulted in more research in intelligent surveillance. It has a wide range of applications, such as observing people in large waiting rooms, shopping centers, hospitals, eldercare, home-nursing, campuses, or monitoring vehicles inside/outside cities, on highways, bridges, and in tunnels [11].

Currently, there is an increasing desire and need for video surveillance applications to be able to analyze human behavior. Behavior analysis involves the analysis and the recognition of motion patterns to produce a high-level description of actions and interactions among objects [12]. Despite significant research efforts over the past few decades, action recognition remains a highly challenging problem. The difficulties of action recognition come from several aspects [13, 14].

- The human movements are represented in a very high dimensional space. Moreover, interactions among different subjects complicate searching in this space.
- Performing similar or identical activities by different subjects exhibit substantial variations.
- The visual data from traditional video cameras can only capture projective information of the real world, and are sensitive to lighting conditions.

Thus, there is a need for an active surveillance system that can perform real-time observation to detect unusual

behavior by capturing the human actions and actuate the alert system.

IV. PROBLEM DEFINITION

In this project, a fuzzy convolution neural network i.e., a convolution neural network with fuzzy inputs for human action recognition based on the features extracted from motion capture information is built. The tracking information of three human joints (right hand, left hand, and pelvis) by using a Kinect camera during the execution of an action is used to compute four distance measures. The nature and range of variation of these distance measures for each action are used to construct the fuzzy membership functions that can emphasize the discriminative pose associated with each action. The temporal variation of membership values of these fuzzy membership functions is used as the discriminative feature for human action recognition. A convolution neural network capable of recognizing local patterns in input data is used to recognizing human actions.

Figure. 1, represents the block model of the proposed system which includes the major blocks like obtaining the video input, computing the features (a, b, c, and d), feeding the Fuzzy Inference System (FIS), obtaining the output of the Fuzzy Inference System (FIS), making decision, and relaying the decision to the alert system.

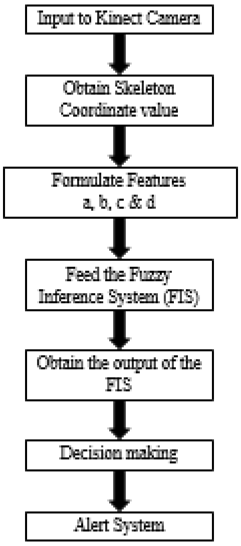
Input to Kinect Camera
Obtain Skeleton Coordinate value
Formulate Features a, b, c & d
Feed the Fuzzy Inference System (FIS)
Obtain the output of the FIS
Decision making
Alert System

Figure 1: Block diagram of the proposed model

Figure. 2 indicates the skeleton model of features formulation, where the skeleton coordinates will be captured from the video surveillance system by using a Kinect camera. These coordinates are used to calculate the features like a, b, c, and d where "a" is the distance between the right and left hand, "b" is the distance between right hand and floor, "c" is the distance between left hand and floor while "d" is the distance between pelvis and floor. Then these features will be trained in the Fuzzy Inference System (FIS). Using these features the human action will be recognized i.e., jumping, bending, punching an object/individual, waving one hand, throwing an object, and kicking an object/individual. On successful identification of the unusual behaviors like fighting or running in a restricted area, the alert system is activated to intimate the concerned individual.

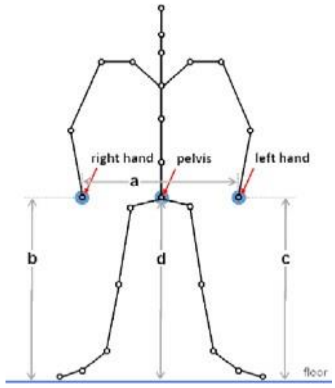

Figure 2: Skeleton model of features (a, b, c, and d) computation

V. PROJECT EXECUTION STEPS

The project is planned to execute with the following steps:

1. Interface Matlab with the Kinect camera. In this following steps are followed:
    - Installation of the Kinect camera.
    - Construction of the Kinect object in Matlab.
    - Obtaining the color and depth image from Kinect camera.
    - Obtain the skeleton coordinates from the depth image.
2. Feature calculation: In this step, the features like a, b, c, and d are calculated.

3. Development of the Fuzzy Inference System (FIS).
4. Designing the Alert System.
5. Integration of all the modules.
6. Testing of the proposed system.

References

[1]. Akila. K and Chitrakala. S, “A Comparative Analysis of Various Representations of Human Action Recognition in a Video”, IJIRCCE, Vol. 2, Issue 1, January 2014.

[2]. Ahmed Taha, Hala H. Zayed, M. E. Khalifa, and El-Sayed M. ElHorbaty, “Exploring Behavior Analysis in Video Surveillance Applications,” In The International Journal of Computer Applications (IJCA), Foundation of Computer Science, New York, USA, Volume 93, Number 14, pp. 22-32. May 2014.

[3]. Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, Juergen Gall, "A Survey on Human Motion Analysis from Depth Data," Lecture Notes in Computer Science, Springer Berlin Heidelberg, Volume 8200, pp 149-187, 2013.

[4]. Lulu Chen, Hong Wei, James Ferryman, “A survey of human motion analysis using depth imagery,” In Pattern Recognition Letters, Elsevier Science Inc., Volume 34, Issue 15, pp. 1995-2006, November 2013.

[5]. Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, “Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group,” In Proceedings of the International IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, Ohio, USA, pp.588-595, June 2014.

[6]. Ahmad Jalal, Shaharyar Kamal, and Daijin Kim, “A Depth Video Sensor-Based Life-Logging Human Activity Recognition System for Elderly Care in Smart Indoor Environments,” In the International Journal of Sensors, Volume 14, Number 7, pp. 11735-11759, July 2014.

[7]. Samuele Gasparrini, Enea Cippitelli, Susanna Spinsante, and Ennio Gambi, “A Depth-Based Fall Detection System Using a Kinect Sensor,” the International Journal of Sensors, Volume 14, Issue 2, pp. 2756- 2775, February 2014.

[8]. Salah Althloothia, Mohammad H. Mahoora, Xiao Zhanga, Richard M. Voylesb, “Human Activity Recognition using Multi-Features and Multiple Kernel Learning,” In Pattern Recognition Journal, Volume 47, Issue 5, pp. 1800–1812, May 2014.

[9]. Jiang Wang, Zicheng Liu, Ying Wu, Junsong Yuan, “Mining Actionlet Ensemble for Action Recognition with Depth Cameras,” In Proceedings of the International IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island, USA, pp. 1290-1297, June 2012.

[10]. Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy, “Sequence of the Most Informative Joints (SMIJ): A New Representation for Human Skeletal Action Recognition,” In proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, Rhode Island, USA, PP. 8-13, June 2012.

[11]. Kavita V. Bhaltilak, Harleen Kaur, Cherry Khosla, “Human Motion Analysis with the Help of Video Surveillance: A Review,” In the International Journal of Computer Science Engineering and Technology (IJCSET), Volume 4, Issue 9, pp. 245-249, September 2014.

[12]. Chen Change Loy, “Activity Understanding and Unusual Event Detection in Surveillance Videos,” Ph.D. dissertation, Queen Mary University of London, 2010.

[13]. Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, Juergen Gall, "A Survey on Human Motion Analysis from Depth Data," Lecture Notes in Computer Science, Springer Berlin Heidelberg, Volume 8200, pp 149-187, 2013.

[14]. Lulu Chen, Hong Wei, James Ferryman, “A survey of human motion analysis using depth imagery,” In Pattern Recognition Letters, Elsevier Science Inc., Volume 34, Issue 15, pp. 1995-2006, November 2013.