International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Detection and Classification of Human Action Events from Captured Video Streams

Dr. Puttegowda D*[1], Suma A P[2]
*[1]Department of Computer Science Engineering, ATME College of Engineering, Mysuru, Karnataka, India
[2]Department of Electrical and Electronics, Vidya Varadhaka College of Engineering, Mysuru, Karnataka, India

ABSTRACT

Human detection and recognizing their actions from the captured video streams is more difficult and challenging task in the field of image processing. The human action recognition is more complex due to variability in shapes and articulation of human body, motions in the background scene, lighting conditions and occlusion. Human actions are recognized by tracking the selected object over the consecutive frames of gray scale image sequences, initially the background motion of the input video stream is subtracted, and its binary images are constructed, the object which needs to be monitored is selected by enclosing the required pixels within bounding rectangle, by using spatiotemporal interest points (Mo-SIFT). The selected foreground pixels within the bounding rectangle are then trackedusing edge tracking algorithm over the consecutive frames of gray scale images. The features like horizontal stride (HS) and vertical distance (VD) are extracted while tracking and the values of these features from the current frame are subtracted with the previous frame values to know the motion. The obtained results after subtraction are then applied to K-Nearest Neiberhood method to recognizing human action using linear prediction technique. This methodology finds an application where monitoring the human actions is required such as shop surveillance, city surveillance, airports surveillance and other places where security is the prime factor.

Keywords: Background Subtraction, Edge Tracking, Classification, Spatio-Temporal Interest Points.

I. INTRODUCTION

Human motion detection and recognizing their actions from the captured video streams is important in the field of image processing, it is extremely complex task to identify the humans and their several types of actions more precisely and accurately[1,3]. Human action recognition finds an application in field of security and surveillance, like shop surveillance, city surveillance, airports and in other places where the security is the prime factor. The great deal of work has been centered in developing systems that can be trained to alert authorities about individuals whose actions appear questionable, for instance in an airport, a system could be trained to recognize a person bending down to leave some baggage and then walking off leaving it unattended as a cause for concern requires investigation, similarly in the department store a person picking up an article and leaving without paying could be interpreted as a suspicious activity[10]. Thus an intelligent and efficient recognition system should identify the actions to know the suspicious activity, so as at inform the security or police personal to take necessary actions before the subject becomes the real culprit.

In the proposed approach, the input video streams are segmented into frames and background motion is subtracted, binary images are constructed by finding

the difference image, which is obtained by calculating the intensity change in each pixel across the frames between image frame (k) and image frame ($k + 1$). The threshold value (T) is calculated by using mean and standard deviation from the difference image. Each pixel in the difference image is compared with the calculated threshold value (T) to subtract the background motion. After the background motion subtraction the object which needs to be monitored is selected, using spatio-temporal interest points (Mo-SIFT) which reduce the whole video frame from a volume of pixels to compact and descriptive interest points.

The edge tracking algorithms are used to track the selected object and the required features are extracted while tracking. The values of the extracted features from the previous frames are subtracted with the current frame value to know the movements which occurred, in the different parts of the human body while performing the human action. Thus selected threshold value to predict the type of human action using linear prediction operation technique.

The human actions considered for recognition are run, walk, jump, bend, hand wave, two hands wave, side walk, skip and multiple actions like, walk-run, walk-hand wave, run hand wave, walk-bend, walk-jump, walk-bend-run and multiple action (two person). The overview of the recognition process is shown in Figure 1.

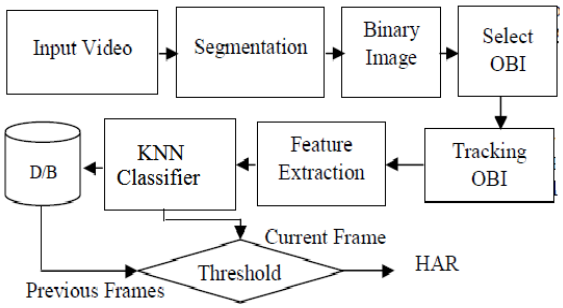

Figure 1. Overview of Human Action Classification Process

II. RELATED WORK

Much of the work has been done in the areas of human detection and human action recognition. Murat Ekinci and Eyiip Gedikip [2] use spatio-temporal jets and silhouette based action recognition techniques, in their approach the gray scale images are used for recognition. The background scene model is statically learned and the pixel having higher redundancy is chosen to have initial background model. The outlines of the foreground object is detected and tracked over successive frames to identify the actions. Nazh Ikizlerand, Pinar Duygulu [4, 5] uses bag of visual rectangles to recognize human actions, in their approach the captured video streams are converted into gray scale images, and its background motion is subtracted using adaptive background subtraction techniques. Histograms of oriented gradients (bag-of-visual-words) is used to represent the selected object as a distribution of oriented rectangular patches and by knowing the orientations of these patches the human actions are recognized. Chunfeng Yuan, Weiming Hu, XiLi, Stephen Maybankand, Guan Luo [6] discuss about human action recognition using log-Euclidean Riemannian metric and histograms of oriented gradients, in their approach, Dollár et al..s detector is used to detect cuboids from each frame.

The proposed approach differs in choice of subtracting the background motion, selecting the object of interest (OBI), tracking the selected object over the successive frames to extract the features, use of linear prediction technique to identify the type of human action based on the chosen threshold.

This paper is structured as follows: Section 3: Classification of the human actions - Proposed Approach Section 4: Presents the experiments and results.

III. PROPOSED APPROACH

The proposed approach has three components say, preprocessing, machine learning and post-processing as in Figure. 2.

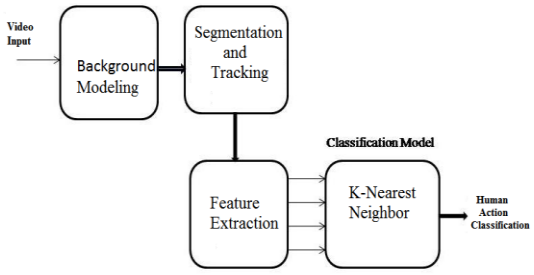

Figure 2. Block Diagram - Proposed Approach

3.1. Pre-Processing

The initial step in human action recognition is preprocessing, which is used to convert the captured coloured video streams into gray scale images, perform background subtraction and constructs binary images for each segmented frames of the captured video stream. The rgb2gray converter is used to convert [RGB] coloured video streams into gray scale images as shown in Eq-1. Let {X1, X1, X3..XN } represents the (N) frames of the segmented video and these frames are converted into gray scale as follows:

$$gray\ (X\ (p_k)) = rgb2gray(X\ (p_k)\ (i, j)) \quad (1)$$

Where k=1, 2, 3...N represents (N) frames and (i, j) indicates the row and column of the selected image frame,and gray (X (p k)) represents the gray scale image ofthe selected frame. After converting each pixel of an image into gray scale its background motion is subtracted, and binary image is constructed by finding the difference image, which is obtained by calculating the intensity change in each pixel across the frames between image frame *k* and image frame *k* + 1.

$$DiffImage\ (i, j) = |\ I_k\ (i, j) - I_{k+1}(i, j)\ | \quad (2)$$

Where 1 ≤ i ≥ N; 1≤ j ≥ M, I^K is the kth image frame and I^{k+1} is the k + 1th image frame. *DiffImage* is the difference image between I^k and I^{k+1}. M and N is the width and length of the image respectively. The statistic characteristics of the difference image are presented by its mean (μ) and standard deviation (δ), for each image point (i, j) of the difference image, the mean intensity *M A* (i, *j*) and the standard deviation *SD A* (i, *j*) are calculated as follows:

$$\mu = \frac{1}{M*N}\sum_{i=1}^{N}\sum_{j=1}^{M} DiffImage(i,j) \quad (3)$$

$$\delta = \frac{\sqrt{\sum_{i-1}^{N}\sum_{j-1}^{M}(DiffImage(i,j)-\mu)^2}}{\sqrt{M-1}\sqrt{N-1}} \quad (4)$$

Where 1 ≤ i ≥ N; 1≤ j ≥ M and *DiffImage* is the difference image. Applying the threshold $T = \mu + 2\,\delta$ *on* each pixel of the difference image, we get the binary motion image as follows BIp:

$$BI_p\ (i, j) = \begin{cases} 1, \text{ if } DiffImage\ (i, j) > T \\ 0, \text{ Otherwise} \end{cases} \quad (5)$$

3.2 Machine Learning

In machine learning algorithms we proceed to select the features which are most useful in recognizing the actions. The optimal solution is the exhaustive search amongst the available features and their combinations, since this approach becomes too complex and time consuming process as the number of features increases. Hence we propose to track only the required features. Initially, after the background subtraction, spatio-temporal interest points (Mo-SIFT) is used to select the object of interest (OBI), by enclosing the required pixels which needs to be tracked within the bounding rectangle. Selecting or detecting sufficient number of interest points containing necessary information to recognize human actions will reduce the whole video frame from a volume of pixels to compact and descriptive interest points. This reduction in volume of pixels is required because most of the human actions are recognized by tracking only the required parts of the human body such as human head, hands and legs instead tracking the whole body. After the object of interest is selected we use edge tracking algorithm to track the selected pixels and to extract the required features. The general tracking method can be broadly classified into two categories: region tracking methods and edge tracking methods [7]. A region tracking method identifies a region of the image, for which it uses similarity measure to decide on the best matching region in the next image of a sequence. The

region is taken to contain some object of interest, with the boundary often being a bounding box, or simple polygon. Edge tracking methods attempt to follow edges, or locations of high luminance or color change, through an image. The edges tracked are usually boundaries of objects of interest within an image sequence.

In our approach we use edge tracking algorithm where the human body which needs to be monitored is enclosed with the bounding rectangle and the edges are tracked over the consecutive frames of gray scale image sequences and to extract the required features like horizontal stride (HS) and vertical distance (VD).

3.3 Post-Processing

The post-processing is the final step in human action recognition. Inputs to this block are the features such as horizontal stride (HS) and vertical distance (VD) which are extracted in the machine learning process. The extracted feature values of the current frame are compared with values of the previous frames to detect the motion in the parts of the human body, this involves in subtracting the feature values α1 and α2 [representing horizontal stride and vertical distance respectively of the previous frame], from the features of the current frame λ1 and λ2 to obtain Δ1 and Δ2 containing the difference values of each pixels which are obtained after the subtraction.

Let λ1 and λ2 represents the extracted feature values (HS & VD) from the current frame of the walking subject, and α1 and α2 represents the feature values obtained from the previous frames, thus we can represent difference as follows:

$$\Delta 1 = \lambda 1 - \alpha 1$$
$$\Delta 2 = \lambda 2 - \alpha 2 \qquad (6)$$

Where Δ1 and Δ2 holds the results obtained after subtraction. The obtained result Δ1 and Δ2 is then compared with the chosen threshold value δ1 and δ2 as in Figure 3 to predict the type of human actions using linear prediction technique.

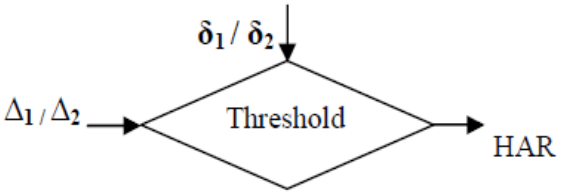

Figure 3. Comparison of threshold values

The human actions considered for recognition are : run, walk, jump, bend, hand wave, two hands wave, side walk, skip and multiple actions like, walk-run, walk-hand wave, run-hand wave, walk-bend, walk-jump, walk-bend-run and multiple action (two person).

IV. EXPERIMENTS AND RESULTS

The proposed approach also recognizes combination of human actions like walk-run, walk-hand wave, run-hand wave, walk-bend, walk- jump, walk-bend-run and multiple actions with higher success rate.

A. Database : KTH

The KTH dataset was used as a standard benchmark for action recognition. It was recorded in four controlled environments with clean background (indoors, outdoors, outdoors with scale variation, outdoors with different clothes.)

The dataset contains about 600 video sequences of 25 subjects performing six categories of actions: boxing, hand clapping, hand waving, jogging, walking, and running. The video resolution is 160x120. we apply exactly the same experimental setting of KTH dataset. Among the 25 persons, we use16 persons (1528 sequences) for training and the other 9 persons (863 sequences) for testing.

Figure 4. Sample frames from the KTH actions sequences.

The confusion matrix for our method is given in table I. Interestingly, the major confusion occurs between jogging and running.

Table 1. Confusion matrix for the KTH actions.

	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	.99	.01	.00	.00	.00	.00
Jogging	.04	.89	.07	.00	.00	.00
Running	.01	.19	.80	.00	.00	.00
Boxing	.00	.00	.00	.97	.00	.03
Waving	.00	.00	.00	.00	.91	.09
Clapping	.00	.00	.00	.05	.00	.95

B. Database : Weizmann dataset

The videos available in Weizmann dataset [8] are considered for recognizing human actions, Totally 100 video samples are considered, the success rate in percentage for different set of actions are shown Table II and in Figure. 5, the proposed approach will identify the human actions correctly for 86 video samples, achieving the overall success rate 86%.

Table 2. Success Rate-Proposed Approach

Sl. No.	Human Actions	Proposed Approach (%)
1	Run	81
2	Walk	100
3	Jump	100
4	Bend	90
5	Hand Wave	89
6	Two Hand Wave	65
7	Side Walk	100
8	Skip	85
9	Walk -Run	63
10	Walk-Hand Wave	100
12	Run-Hand Wave	100
13	Walk-Bend	63
14	Walk-Jump	79
15	Walk-Bend-Run	82

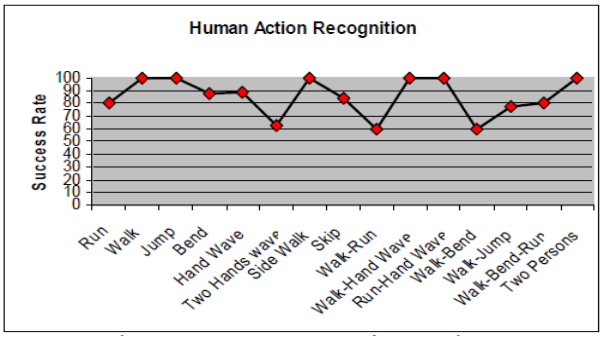

Figure 5. Success rate- Proposed Approach

V. CONCLUSION

Algorithms for detection and classification of human actions from the camera feed are proposed. The proposed work involves identifying the actions by extracting the features from the obtained binary image. The captured videos are pre-processed to subtract the background motion, and the features such as horizontal stride and vertical distance are extracted from the binary images by tracking its motion over the successive frames using machine learning algorithms. In post-processing these extracted features are then compared with the chosen threshold values to identify the human actions. Multiple camera feeds can be integrated together to improve the accuracy of the classification, and as well as to reduce the problems of occlusion, noise, poor lighting condition, contrast and brightness.

VI. REFERENCES

[1]. Puttegowda D and Dr. M C Padma,"A Framework for Event Classification from Video Sequences using Bayesian Neural Network",Communications on Applied Electronics (CAE),Volume: 05,Issue:

02,Pages:1-5,DOI: 10.5120/cae2016652229,ISSN : 2394-4714,May,2016

[2]. Murat EKINCI,Eyup GEDIKLI (2005) Silhouette Based Human Motion and Action Detection and Analysis for Real-Time Automated Video Surveillance. Turk J Elec Engin. Volume 13,No.2.

[3]. Puttegowda D and Dr. M C Padma,"Human Motion Detection and Recognising their Actions from the Video Streams" International Conference on Informatics and Analytics (ICIA'16),Pondicherry,DOI: http://dx.doi.org/10.1145/ 2980258.2980290,August-2016

[4]. Nazh Ikizler and Ponar Duygulu (1999) Human Action Recognition Using Distribution of Oriented Rectangular Patches. Computer vision and pattern recognition (CVPR.05).Volume1,pp 886-893.

[5]. Nazh Ikizler and Ponar Duygulu (2009) Histograms of oriented rectangles: A new pose descriptor for human action recognition. Image and vision computing .Volume 27,Issue 10,pp 1515-1526.

[6]. Chunfeng Yuan,Weiming Hu,Xi Li,Stephen Maybank,Guan Luo (2004) Human Action Recognition under Log-Euclidean Riemannian Metric. Computer Vision .ACCV2009 9th Asian Conference on computer vision.

[7]. Haiying Guan,Ronda Venkateswarlu Adaptive Multimodal Algorithm on Human Detection and Tracking.

[8]. Link to Weizmann Dataset : http://www.wisdom.weizmann.ac.il /~vision/SpaceTime Actions.html.

[9]. R.Venkatesh Babu and R.Hariharan (2009) Image processing,video surveillance,and security related applications using parallel machines. NAL-PD-FS-0916 National Aerospace Laboratories.

[10]. Puttegowda D,Deepak N.A and Rajesh Shukla,"Robust Image Transmission over Nose Channel using Independent Component Analysis",International Journal of Advanced Networking & Application,Volume: 01,Issue: 06,Pages:347-352 (2010).