Silhouette Based Human Motion Detection and Recognising their Actions from the Captured Video Streams

Deepak. N. A
Flosolver Division, National Aerospace Laboratories (NAL), Bangalore, India.
Email: deepak.n.ananth@gmail.com

Dr.U. N. Sinha
Head, Flosolver Division, National Aerospace Laboratories (NAL), Bangalore, India
Email: sinhaun47@gmail.com

---ABSTRACT---

Human detection and recognizing their actions from the captured video streams is more complex 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 spatio-temporal interest points (Mo-SIFT). The selected foreground pixels within the bounding rectangle are then tracked using 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 compared with the selected threshold value to predict the type of 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, Linear Prediction, Occlusion, Spatio-Temporal Interest Points (Mo-SIFT), Surveillance, Threshold**.

Date of Submission: 19 December 2010 Date of Acceptance: 29 January 2011

1.0 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. 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. 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 obtained result after subtraction is then compared with the 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 Fig. 1.

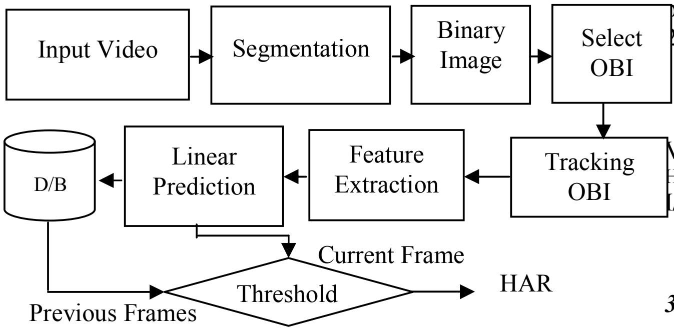


Fig. 1: Overview of Human Action Recognition Process

2.0 Related Work

Much of the work has been done in the areas of human detection and human action recognition. Murat Ekinci and Eyiip Gedikip [1] 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 [2, 3] 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 [4] 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. Then the descriptors are used to extract the features (bag-of-visual-words) by using the k-mean clustering method. Histograms of bag-of-visual-words are then classified according to histograms matching between the test video and training video sequence. EMD matching techniques are then used for matching each video pair obtained from training and testing samples to identify the human actions.

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: Recognizes the human actions - Proposed Approach Section 4: Presents the experiments and results.

3.0. Recognizing Human Actions – Proposed Approach

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

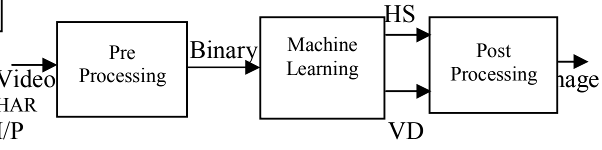


Fig. 2: Block Diagram - Proposed Approach

3.1. Pre-Processing

The initial step in human action recognition is pre-processing, 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 $\{X_1, X_1, X_3 \ldots X_N\}$ represents the (N) frames of the segmented video and these frames are converted into gray scale as follows:

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

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 of the 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.

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

Where $1 \leq i \geq N$; $1 \leq j \geq M$, 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 $1 \leq i \geq N$; $1 \leq j \geq M$ 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 BI_p:

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

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 [5]. 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 α_1 and α_2 [representing horizontal stride and vertical distance respectively of the previous frame], from the features of the current frame λ_1 and λ_2 to obtain Δ_1 and Δ_2 containing the difference values of each pixels which are obtained after the subtraction.

Let λ_1 and λ_2 represents the extracted feature values (HS & VD) from the current frame of the walking subject, and α_1 and α_2 represents the feature values obtained from the previous frames, thus we can represent difference as follows:

$$\Delta_1 = \lambda_1 - \alpha_1$$
$$\Delta_2 = \lambda_2 - \alpha_2 \quad (6)$$

Where Δ_1 and Δ_2 holds the results obtained after subtraction. The obtained result Δ_1 and Δ_2 is then compared with the chosen threshold value δ_1 and δ_2 as in Fig 3 to predict the type of human actions using linear prediction technique.

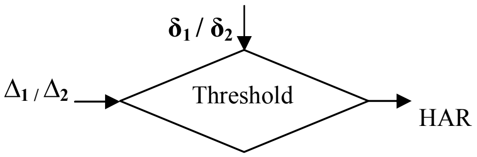


Fig. 3: Shows the result of comparison

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).

3.3.1 Linear Prediction Operation Technique

Linear prediction is a mathematical operation, where the future values are estimated as a linear function of previous frame samples. Linear prediction operation is closely linked to modelling of vocal track systems and relies upon the fact that speech samples may be predicted by the linear combination of previous samples. The number of previous samples used for prediction is known as the order of the prediction which is used to minimize the prediction error.

In the proposed approach linear prediction technique is used to predict the type of human actions. The type of action is predicted by comparing the values of the extracted features from the previous frames P (z) with the feature values of the current frame C (z) to predict the type of human action S (z) as in Fig. 4.

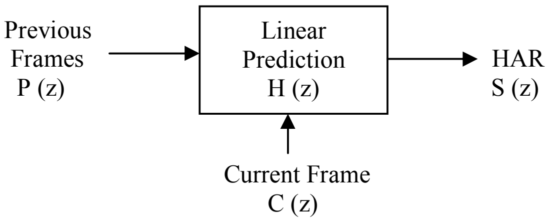


Fig. 4: Human Action Recognition by Prediction

Where prediction function H (z) is represented as:

$$H(z)= \frac{S(z)}{P(z)\,C(z)} \qquad (7)$$

Here we compare the feature values extracted from the previous frames with the current frame values to recognize human actions.

This involves in finding prediction model represented as:

$$S_n = \sum_{i=1}^{P} P_i X(n-i) + C_n \qquad (8)$$

Where Sn is the predicted value, X (n − i) the previous observed values, and (P_i) the predictor coefficients. The error generated by this estimate is

$$E(n) = X(n) – S(n) \qquad (9)$$

Where x (n) is the original value, S (n) is the predicted value. The differences are found in the way the parameters (P_i) are chosen.

4.0 Experiments & Results

Several methods exists to detect and recognize the human actions from the captured video streams, here we compare the results of the other algorithms like Spatio-Temporal Jets (ST-Jets), Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradient (HOG) and Log-Euclidean Riemannian Metric algorithms (LERM) [1,2, 3 and 4] with the results of the proposed method, it is found that, for actions like jump, walk, run and hand wave, the success rate of the proposed method is much better than other algorithms used for comparison, and for bend action, other algorithms give better results than the proposed approach. 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.

The videos available in Weizmann dataset [6] are considered for recognizing human actions, Totally 100 video samples are considered, the success rate in percentage for different set of actions are shown in Fig. 5, the proposed approach will identify the human actions correctly for 86 video samples, achieving the overall success rate 86% Fig. 6 shows the comparison of result obtained by the proposed approach with the results of other algorithms, it is observed that, the proposed approach gives better and more efficient results for most of the actions of Table1.

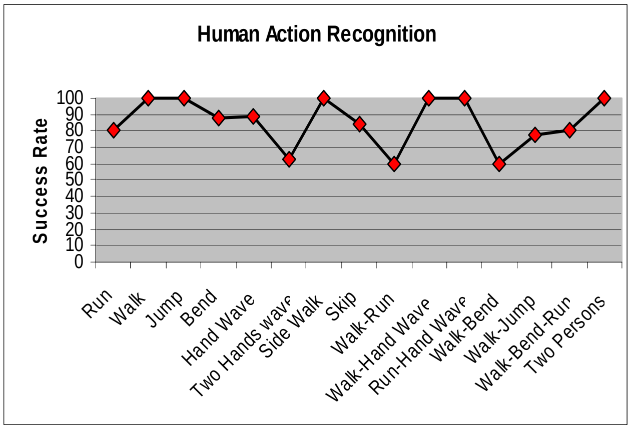


Fig. 5: Success rate – Proposed Approach

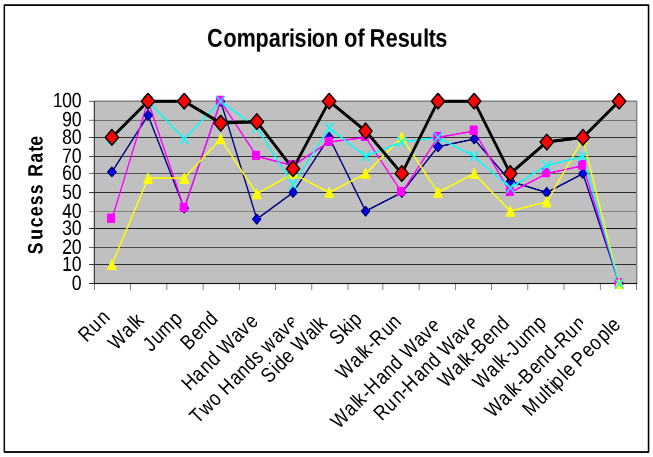


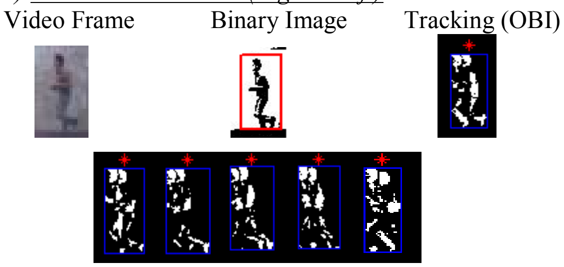
Fig. 6: Comparison of results

Table I : Success in percentage - Comparisons

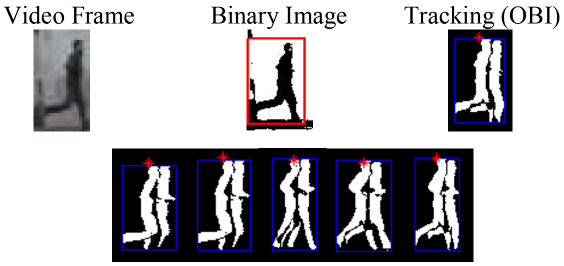
Dataset (Human Actions)	ST-Jets (%)	SIFT (%)	HoG (%)	LERM (%)	Proposed Approach (%)
Run	61	35	10	79	80
Walk	92	100	58	100	100
Jump	41	41	58	79	100
Bend	100	100	79	100	88
Hand Wave	35	70	49	85	89
Two Hands wave	50	65	60	54	63
Side Walk	80	78	50	85	100
Skip	40	80	60	70	84
Walk-Run	50	50	80	78	60
Walk-Hand Wave	75	80	50	80	100
Run-Hand Wave	79	84	60	70	100
Walk-Bend	56	50	40	53	60
Walk-Jump	50	60	45	65	78
Walk-Bend-Run	60	65	79	70	80
Multiple People : Two Persons	--	--	--	--	100

4.1 Results (Screen Shots)

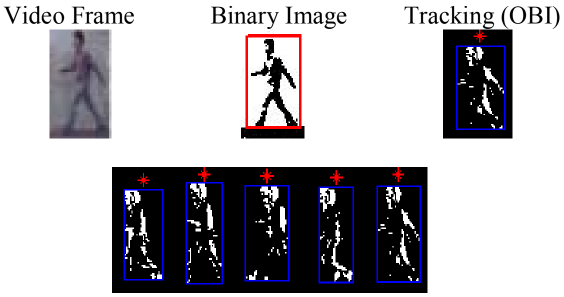
a) Human Action – Run (Right – Left)


Video Frame Binary Image Tracking (OBI)

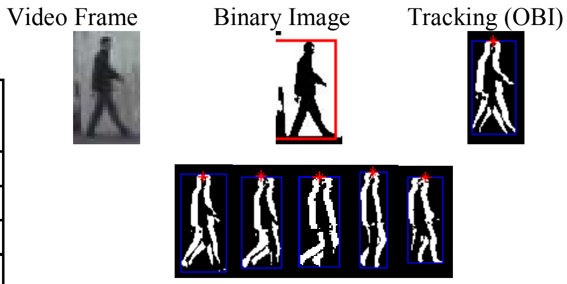
b) Human Action – Run (Left- Right)


Video Frame Binary Image Tracking (OBI)

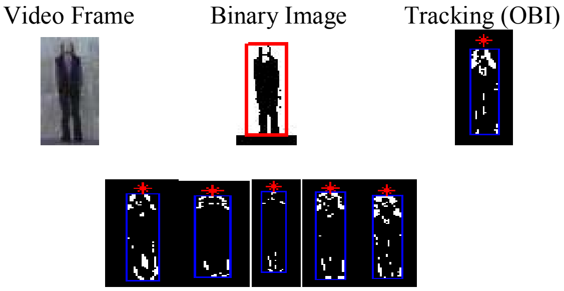
c) Human Action – Walk (Right-Left)


Video Frame Binary Image Tracking (OBI)

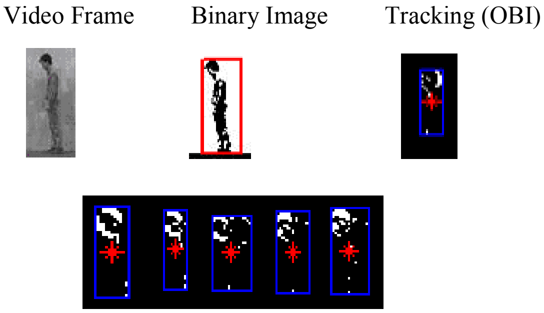
d) Human Action – Walk (Left- Right)


Video Frame Binary Image Tracking (OBI)

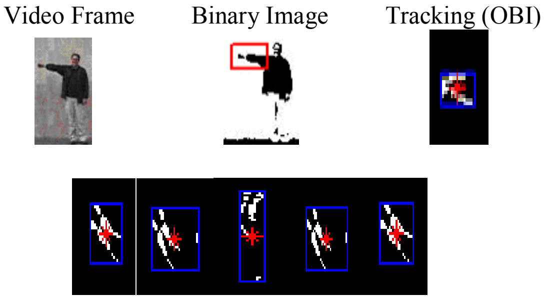
e) Human Action – Jump


Video Frame Binary Image Tracking (OBI)

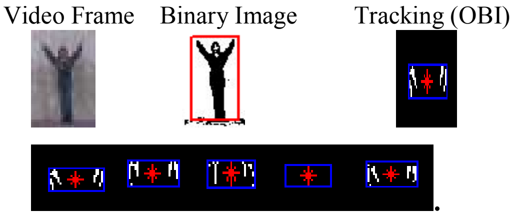
f) Human Action – Bend


Video Frame Binary Image Tracking (OBI)

g) Human Action – Hand Wave

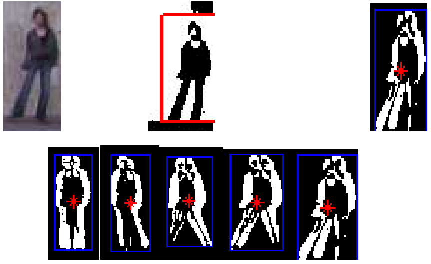

Video Frame Binary Image Tracking (OBI)

h) Human Action – Two-Hands-Wave


Video Frame Binary Image Tracking (OBI)

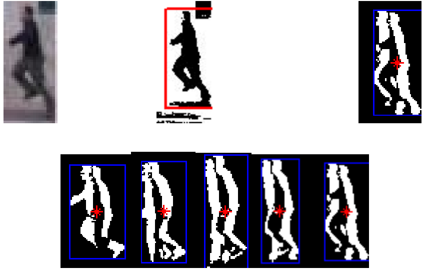
i) Human Action – Side Walk

Video Frame Binary Image Tracking (OBI)



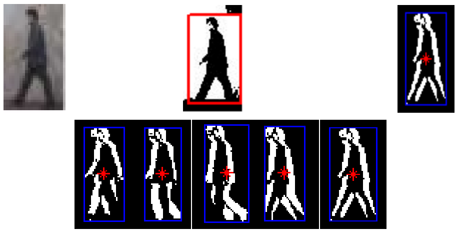
j) Human Action – Skip

Video Frame Binary Image Tracking (OBI)



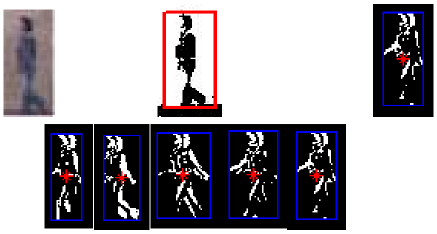
k) Human Action – Walk - Run

Video Frame Binary Image Tracking (OBI)



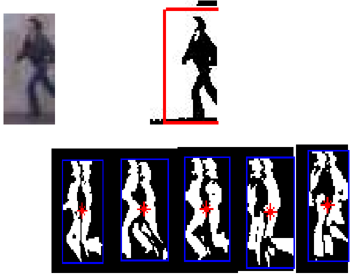
l) Human Action – Walk-Hand wave

Video Frame Binary Image Tracking (OBI)



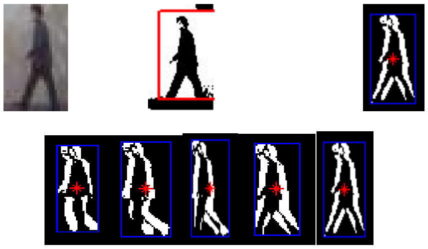
m) Human Action – Run -Hand wave

Video Frame Binary Image Tracking (OBI)



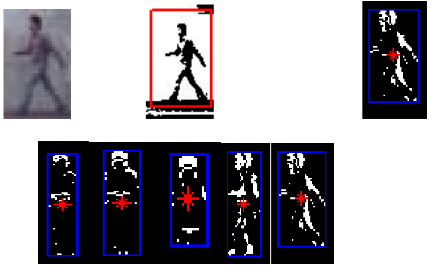
n) Human Action – Walk-Bend

Video Frame Binary Image Tracking (OBI)



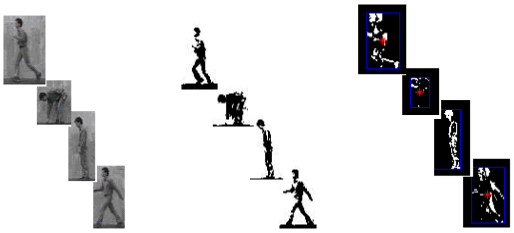
o) Human Action – Walk-Jump

Video Frame Binary Image Tracking (OBI)



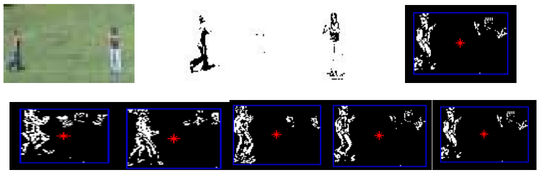
p) Human Action – Walk-Bend-Run

Video Frame Binary Image Tracking (OBI)



q) Multiple Objects (Human's) - Two Persons

Video Frame Binary Image Tracking (OBI)



5.0 Parallel Implementation

The high-performance computing during recent years, occupies an important role in real-time surveillance applications, where large-scale image and video processing is involved. Because it is critical to complete the analysis in a short time for its effectiveness recognizing the human actions from captured video streams is much more computationally intensive. The time taken to access, decode and analyse 1-day video streams from one camera feed is 34 hours in desktop machine and 26 minutes in NAL Flosolver Mk-8 parallel machine [7] and this computation time will be even lesser in new 128 bit Flosolver's parallel machine. Hence this work also envisages the implementation of proposed approach on Flosolver's parallel machine.

6.0 Conclusion and Future Enhancement

Algorithms for tracking and identification 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.

Acknowledgement

I would like to thank Dr.R.Hariharan, for accepting to supervise my research work. He helped in a broad range of issues from giving me direction, helping to find solutions to problems, outlining requirements and always having time to help me.

References

[1]. 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.

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

[3]. Nazh Ikizler and Pınar 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.

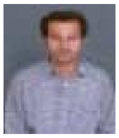
[4]. Chunfeng Yuan, Weiming Hu, Xi Li, Stephen Maybank, Guan Luo (2004) Human Action Recognition under Log-Euclidean Riemannian Metric. *Computer Vision –ACCV2009 9th Asian Conference on computer vision.*

[5]. Haiying Guan, *Ronda Venkateswarlu* Adaptive Multi-modal Algorithm on Human Detection and Tracking.

[6]. Link to Weizmann Dataset : http://www.wisdom.weizmann.ac.il /~vision/Space Time Actions.html.

[7]. 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.*

Authors Biography



Mr. Deepak N. A. working as Assistant Professor in the department of Computer Science & Engineering, Ghousia College of Engineering Ramanagaram, Karnataka, currently pursuing research work on image processing at National Aerospace Laboratories (NAL) research centre, Bangalore, affiliated to University of Mysore, and has nine years of teaching experience, and has published one national paper, one international paper and two international journals in the field of image processing and data mining.

Dr. U N Sinha, the scientist, is very well known in the NAL, CSIR and the Indian scientific circles for his breadth and depth of knowledge in mathematics, fluid dynamics thermodynamics, parallel computing, atmospheric science and Sanskrit. Dr. U N Sinha obtained his engineering degree engineering in 1967 and his PhD from IIT Kanpur in 1976. In 1986 Flosolver project was born. It is creditable that Dr. Sinha and his team developed Flosolver Mk1, India's first parallel computer and in 1993 Dr.U N Sinha and his team was first to complete the parallelization taking up of DST's project to parallelize a global weather prediction model which was being used by NCMRWF for operational forecasts. Apart from his scientific achievements, Dr. Sinha has many other accomplishments to his credit: building a very good book collection in the library, teaching a very large number of students, taking good care of his colleagues.