Human Activity Recognition System

A dissertation submitted in partial fulfillment of the requirement for the Master of Technology in Computer Science degree of the Indian Statistical Institute

By

A. AravindKumar Yadav

Under the supervision of

Prof **Dipti Prasad Mukherjee, Ph D, FIE**

Electronics and Communication Sciences Unit

Computer and Communication Sciences Division

INDIAN STATISTICAL INSTITUTE

203, Barrack pore trunk road

Kolakata-700108

2008-2010

Chapter 1

Introduction

The main objective of this project is to identify human activities mainly running and walking. In this report, we present a method for human activity Recognition in video. Human activity recognition from video streams has applications in choreography, sports, security surveillance, content based retrieval motion analysis, virtual reality interfaces, robot navigation and recognition, video indexing, browsing, HCI etc. The complexity involved in human activity varies from simple hand gestures to many body parts a lot.

We build an approach to analyze the periodicity in human actions. Here we have considered only two types of activities. They are running and walking. Our method exploits the correlation between the frames for 3 seconds length of time, then identifies the activity periodicity

The system consists of following stages:

i) Tracking
ii) Feature Extraction
iii) Classification.

Of the above stages, before we track an object, first we are manually selecting the region of interest by clicking on a pixel. We consider a region of M X N sized window around the pixel of interest making pixel of interest as the center of the window. We find the best match of the region in next frame and we track an object till all the frames, will give the trajectory of each and every bounding box that contains our pixel of interest. In this report we mainly presented an approach using the waveforms of a tracking pixel i.e. trajectory of every pixel of interest and finding out their properties using signal processing techniques, extracting features from them and train the system with support vector machines and classify the new videos.

Several approaches for activity recognition have been reported in the literature [l]. Previous approaches employed methods such as time-delay neural

networks [5], Hidden Markov Models [3][4] or dynamic time warping [2] to recognize hand gestures and articulated human activity.

1.1 Motivation for Human Activity Recognition

Human action recognition is a very important component of visual surveillance systems for event based analysis of surveillance videos. Visual surveillance systems play a very crucial role in the circumstances where continuous patrolling by human guards is not possible like international border patrolling, nuclear reactors etc. Demand for automatic surveillance systems in civilian applications like monitoring a parking lot, shopping complexes etc. is also increasing heavily. It is difficult and manpower intensive to monitor the data collected from various cameras continuously and this gives rise to the necessity for automatic understanding of human actions and building a higher level knowledge of the events occurring in the scene by the computer vision system.

Analysis in surveillance scenarios often requires the detection of abnormal human actions. Most of the normal human activities are periodic like walking, running etc. Lack of periodicity is therefore an important cue of an activity being deviant from the normal. Consider for example a typical event of surveillance interest: exchange of brief cases by two agents. The scene essentially consists of an agent walking across the scene who then bends to lift up or leave the briefcase. This event can be described as concatenation of walk-bend-walk actions, where bend is deviant from normal behavior. However abnormal events and therefore abnormal human activities are context dependent and may vary for different situations. For example, in a shopping mall where people normally walk from one counter to another, running could be defined as an abnormal action and could be an event of interest for surveillance purposes. This calls for a need of unified framework for detecting and recognizing both periodic and non periodic human actions.

Recognition of human movements has also been exploited to a large extent for animation like avatar control, for giving gesture based commands to virtual reality interfaces, human computer interactions in smart room like environments etc. Content based video retrieval, indexing and searching is also

becoming popular these days with the concepts like Video Google. These systems require cognitive vision techniques for analyzing videos which in real life scenarios mostly converges to analyzing human actions in the videos. Video annotation of sports videos is an excellent example of this category where complex human sport actions are required to be classified. A good discussion on the promising application scenarios and the suitable approaches in these scenarios can be obtained in [4]. The wide scale applications of human activity analysis and various challenges involved at different stages in building this system makes it a demanding area of research in computer vision.

1.2 Organization

In the second section we mainly talk about the related work done in human activity recognition system and our proposed way of solving the problem.

In the third section we present the tracking method and learning methods in detail.

In the fourth section we give the results, that also include tracking results.

In the fifth section we conclude and present the future work.

Chapter 2

2.1 Related Work

Human activity recognition is a long-studied problem in the field of computer vision. Since human activities are highly spatial and temporally structured entities, a large portion of the existing approaches are based on graphical models, such as Finite State Machines (FSM), Hidden Markov Models (HMM), Context-Free Grammar (CFG) and Dynamic Bayesian Networks (DBN). Mahajan et.al., [7] have proposed a model of multi-layered FSMs, which is built on top of spatiotemporal video features and primitive object detection output. The Hidden Markov-Models and its variations are very popular approaches to activity recognition. One representative work is by Hongeng and Nevatia [8]. In their system, a semi hidden Markov model is constructed by using the shape and motion features of tracked objects. Laxton et.al., [9] proposed a Dynamic Bayesian Network (DBN) structure which incorporates partially ordered sub-actions, a hierarchical action representation for building complex actions and an approximate Viterbi inference algorithm.

Since most graphical models are only assumed to handle sequential events, they are not capable of capturing activities with parallel actions. To address this problem, several contributions have been developed. Pinhanez [10] proposed one of the first works in modeling events with durations by incorporating Allen's interval algebra. Shi et.al., [11] proposed Propagation Networks (P-Nets), which model sequential activities with concurrent primitives. Different from many graphical models where primitives are considered instantaneous, primitives in these methods are assumed to have temporal durations such that interval logic can be applied.

In addition to the above mentioned graphical models, other state-based approaches have also been developed. Filipovyh and Ribeiro [12] presented a probabilistic model to capture human-object primitive interactions. In their framework, both static and dynamic appearances of the actor and the target object are encoded in a joint distribution. The intrinsic spatiotemporal configurations

between actor and objects are also modeled. Other approaches have also been pursued. Rao et.al. [13] presented a rank theory for matching trajectories. Activity trajectories are matched by analyzing the rank of their observation matrix. Boiman and Irani [14] have applied the spatiotemporal interest-point descriptor to detect irregular activities by comparing each descriptor with its neighbors in the spatiotemporal dimensions. Wong et.al., [15] have extended the probabilistic latent semantic analysis (pLSA) model to incorporate both visual parts and the structural information between visual parts to classify activities in videos. The major limitation here is that descriptor-based methods are not able to either capture the temporal order of the events or handle overlapping scenarios.

2.2 Proposed Method

In our method we considered a rectangular block with center as the tracking pixel. We manually select the pixels of interest that is the tracking pixel. To find the region or block of interest in the next frames by simply estimating the correlation coefficient in the bigger region may be twice the size of the block of interest in the original frame and it can be varied. The highest correlation value corresponding to the block is considered to be the best match. Then we track region of interest in all the frames and take the path or trajectory it traversed with its center. Now divide the image into 5 horizontal blocks and consider all the pixels(of interest) trajectory in each block as the signals and sum it up. Now take the direct cosine transform of all the pixels in 5 blocks and taking the features from them and performing leave one out using Support Vector Machines to classify.

Chapter 3

3.1 Methodology

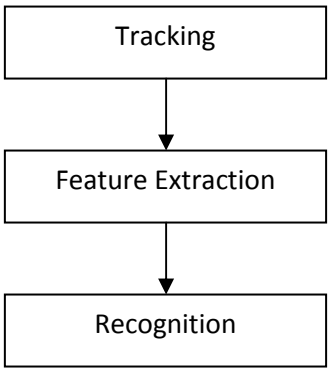


Fig (i) flow chart of our activity recognition system

Figure (i) shows typically the steps involved in Activity Recognition system. Each of the blocks is an area of research in own right. In this work, we mainly concentrated on the Second block i.e. feature extraction for action recognition

Human body has a large number of degrees of freedom. Modeling structural and dynamic features for action recognition of such a complex object is a tough task. Analyzing human action is particularly challenging owing to the complex non rigid and self occluding nature of the articulated human motion.

Implementing real life activity recognition system is a difficult task considering the challenges at each stage of the system like background clutter, dynamic illumination changes, camera movements etc. in the background subtraction stage, partial occlusions in the tracking and feature extraction stages. The performance of the recognition stage depends on these previous stages and also on the choice of features for action representation. The action classification problem is characterized by large intra class variability introduced by various sources like the changes in camera viewpoint, body shapes and sizes of different actors, different dressing styles, changes in execution rate of activity, individual styles of actors etc.

3.2 Background Removal

We can use a GMM (Gaussian Mixture Model) background model similar to the one described in [16] .GMM is very popular in image sequence analysis due to some advantages namely, adaptivity, time-efficiency, robustness etc. The recent history of each pixel X_1, X_2,..X_n is modeled by a mixture of Gaussian distributions. The probability of observing the current pixel value is

$$P(X_t) = \sum_{i=1}^{K} w_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \qquad \text{...3 (a)}$$

where K is the number of the distributions, $w_{i,t}$ is an estimate of the weight (what portion of the data is accounted for by this Gaussian) of the i^{th} Gaussian in the mixture at time t, $\mu_{i,t}$ is the mean value of the i^{th} Gaussian of the mixture at time t, $\sum_{i,t}$ is the covariance matrix of the i^{th} Gaussian of the mixture at time t, where η is a

Gaussian probability density function

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-1/2(X_t - \mu_t)^T \Sigma^{-1}(X_t - \mu_t)} \qquad \text{...3 (b)}$$

K is determined by available memory (we use K between used 3 to 5). The covariance matrix is assumed to be of the form $\sum_{k,t} = \sigma_k^2 I$ that is the red, green, blue pixel values are independent and have the same variance. Every time a new pixel value X_t occurs it is checked for its belongingness to one of the K distributions. If it lies within 3 standard deviations of a distribution it matches to that distribution.

Here, the prior weights of K distributions at time t,$w_{k,t}$ are adjusted as follows

$$w_{k,t} = (1 - \alpha) w_{k,t-1} + \alpha(M_{k,t}) \qquad \text{...3 (c)}$$

where α is the learning rate, $M_{k,t}$ is for the model which matched and is 0 for the remaining models. Then the weights are normalized. The μ,σ parameters for the unmatched distributions remain the same. The parameters of the distribution which matches the new observation are updated as follows

$$\mu_t = (1-\rho)\mu_{t-1} + \rho(X_t) \qquad \text{...3 (d)}$$

$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + \rho\left(X_t - \mu_t\right)^T\left(X_t - \mu_t\right) \qquad \text{...3 (e)}$$

where

$$\rho = \alpha\eta\left(X_t|\mu_k,\sigma_k\right) \qquad \text{...3 (f)}$$

Gaussians are ordered by the value of ω/ σ. Then the first B distributions are chosen as background model, where

$$B = \arg\min_b \left(\sum_{k=1}^{b} w_k\right) \qquad \text{...3 (g)}$$

where T is a threshold set by the user. The following parameter values were empirically observed to provide the best performance α=0.1variance values for GMM initialization σ=3 and minimum weightage for a distribution to remain in the GMM as 0.05.

We can also use a simple shadow detection method, similar to the one given in [17]. We can make an assumption that the shadows don't change the hue significantly, which is valid over most general scenarios, except in case of hard shadows. In such cases sophisticated shadow removal techniques which use static edge-texture information in foreground detected regions to mask of the shadowed parts, can be used. We remove the shadows using two properties: shadows have lower brightness than the background i.e shadows have photometric gain with

respect to background image which is less than 1 and which is reasonably constant over the shadow region. And shadows do not change the chromaticity to a large extent. Shadow detection is performed only for the pixels previously detected as foreground. Shadow mask is defined as,

$$S_t(x,y)=1 if\ photometric\ gain = I_t(x,y)/B_t(x,y)<1 \quad \text{...3 (h)}$$

And

$$|I_t(x,y) - B_t(x,y)| < threshold1 \quad \text{...3 (i)}$$

And

Variance (photometric gain) < threshold2

$$S_t(x,y) = 0\ otherwise \quad \text{...3 (j)}$$

We can use the above shadow mask to remove shadow pixels. After shadow removal stage, the occurrences of misclassified pixels as the foreground are further removed by spatial voting. It scans 5X5 neighborhood of every foreground detected pixel and votes this pixel as foreground or background depending on whether more than 50% (voting threshold) of its neighboring pixels are foreground or background respectively. This helps to eliminate local, noise caused aberrations from being classified as foreground. We further perform connected component analysis (CCA) to get rid of other spuriously detected foreground blobs. Since we assume that in the scene at a time single agent is present. In multi agent scenes our approach can be applied by tracking different agents separately and handling one agent at a time.

In order to avoid the above segmentation process we have taken our own live videos mostly that does not have background problem.

3.3 Tracking

Different kinds of algorithms use different criteria for comparison of blocks. One of the algorithms to be used for block matching is called the Full Search or the Exhaustive Search. In this, each block within a given search window

is compared to the current block and the best match is obtained (based on one of the comparison criterion). Although, this algorithm is the best one in terms of the quality of the predicted image and the simplicity of the algorithm, it is very computationally intensive. There are a number of criteria to evaluate the "goodness" of a match and some of them are:

1. Cross Correlation Function
2. Pel Difference Classification (PDC)
3. Mean Absolute Difference
4. Mean Squared Difference
5. Integral Projection

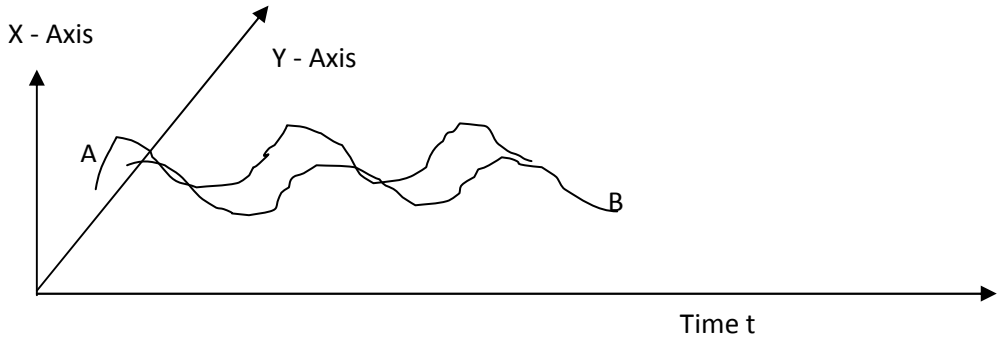
But we have used cross correlation function to find the best matching block in the next frames. We calculate correlation coefficient is computed as

$$r = \sum_m \sum_n (A_{mn} - \mu_A)(B_{mn} - \mu_B) / \sqrt{(\sum_m \sum_n (A_{mn} - \mu_A)^2 (B_{mn} - \mu_B)^2)} \quad \ldots 3 \text{ (k)}$$

where r = correlation coefficient μ_A= mean of matrix A elements, and μ_B = mean of matrix B elements.

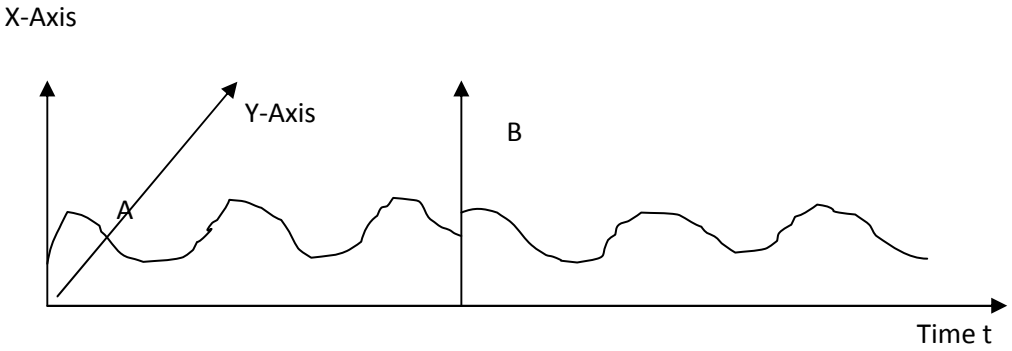
3.4 Learning

Now we divide the whole image into 5 horizontal regions and we take the trajectory of each and every pixel of interest in that horizontal region. Consider the trajectories as signals and combine them one signal after another as shown in the figure below. Let A & B be the trajectories of two pixels or regions of interest.



Fig(ii)

Fig (ii) shows original trajectory of pixels A & B



Fig(iii)

Fig (iii) shows trajectory joined one after another

And after take the combined signal and calculate the direct cosine transform of the signal that yields real values only with one DC component and the other alternating components. And now we will have one such vector of components for each horizontal region that is totally we have 5.Of these above vectors we have taken the first 15 values in each horizontal strips to be the features. We have chosen first 15 because as the variance is maximum for these values and all the remaining have very less variance. Now for each video we have 15 five dimensional points. The system is trained using SVM.

As we have only two types of activities to be identified, the total videos are classified into two. A new video is taken and the above features are extracted that gives us a 15 five dimensional points. These 15 five dimensional

points are classified and can say that a video belong to the class1 or class 2 based on the best match.

CHAPTER 4
4 Results

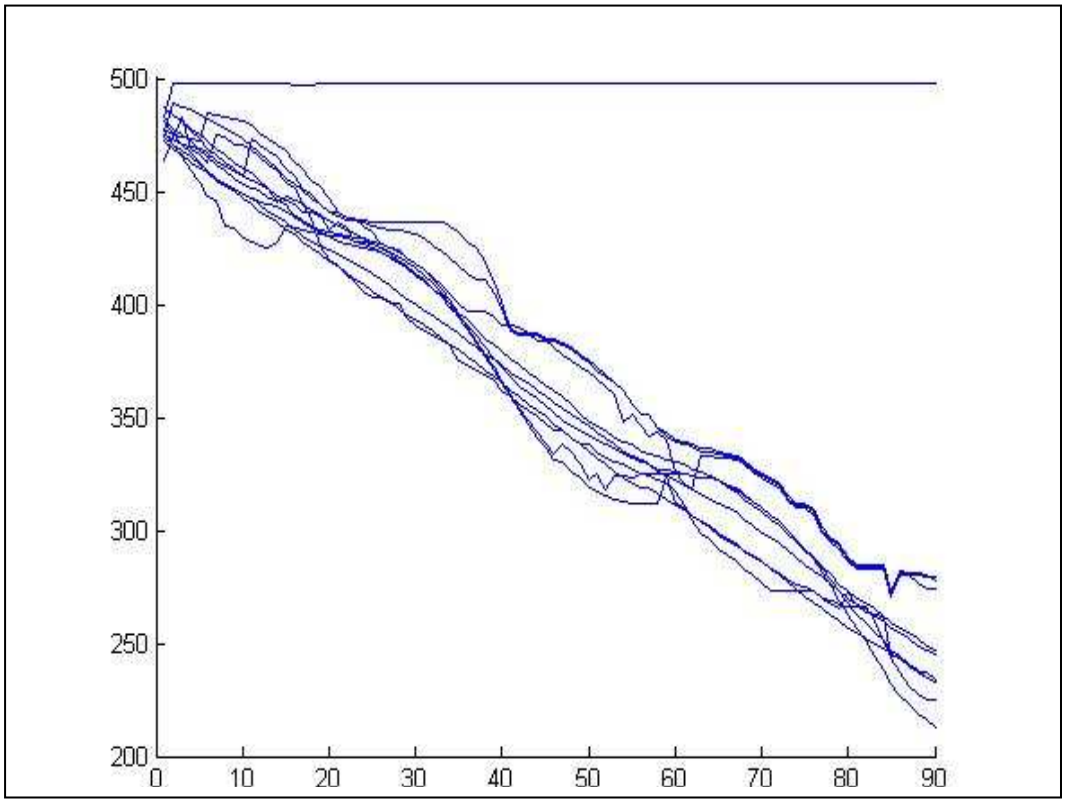
4.1 Experimental Details

We have used Sony Cyber-Shot DSC H20 to shoot the videos.

4.2 Data Details

We have taken 5 Actors performing two activities (i.e., walking and running). We have taken the videos of length 2 – 3 seconds of each activity with 5 such instances.

Y - Axis

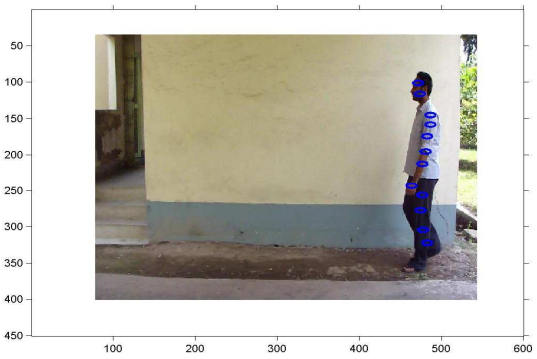


Time t

Fig (iv)

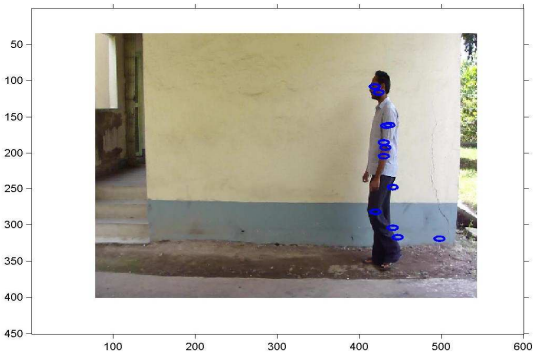
Fig (iv) shows the the trajectory of the pixels of interest with ‘t’ as frame number viewed as 2d graph for walking.

The figures from (v a) –(v g) below shows the tracking pixels from first frame till the last frame, actor is walking from right to left. And the X and Y denote the axis that represents pixel position.

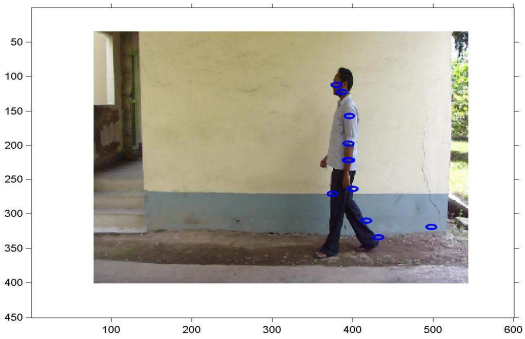


(v. a)

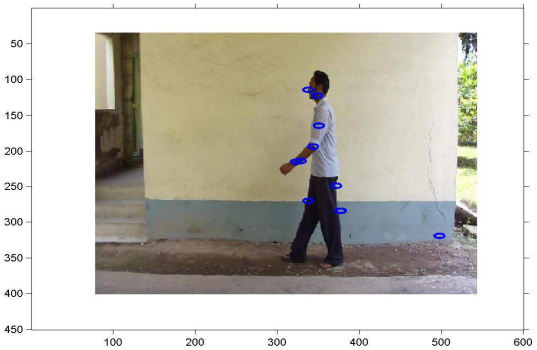
(v. b)



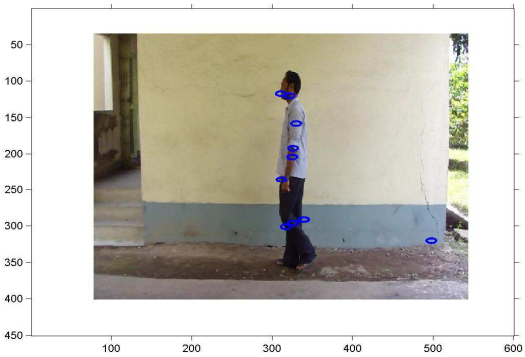
(v. c)



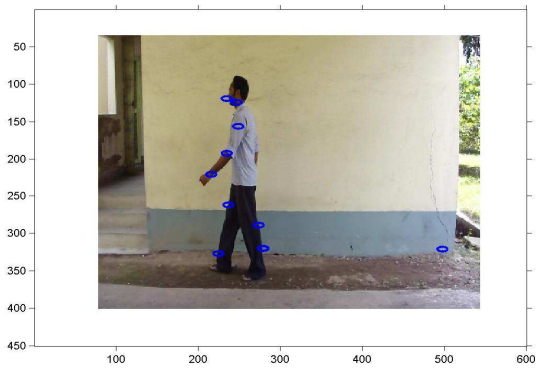
(v. d)



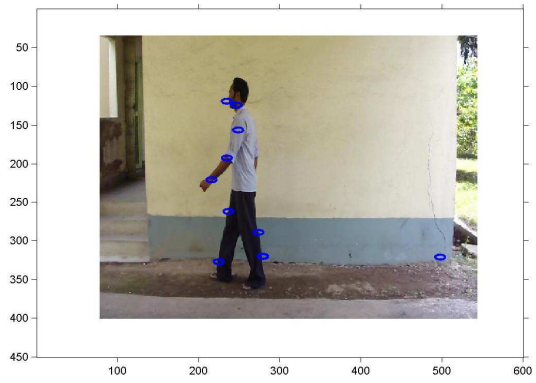
(v. e)



(v. f)

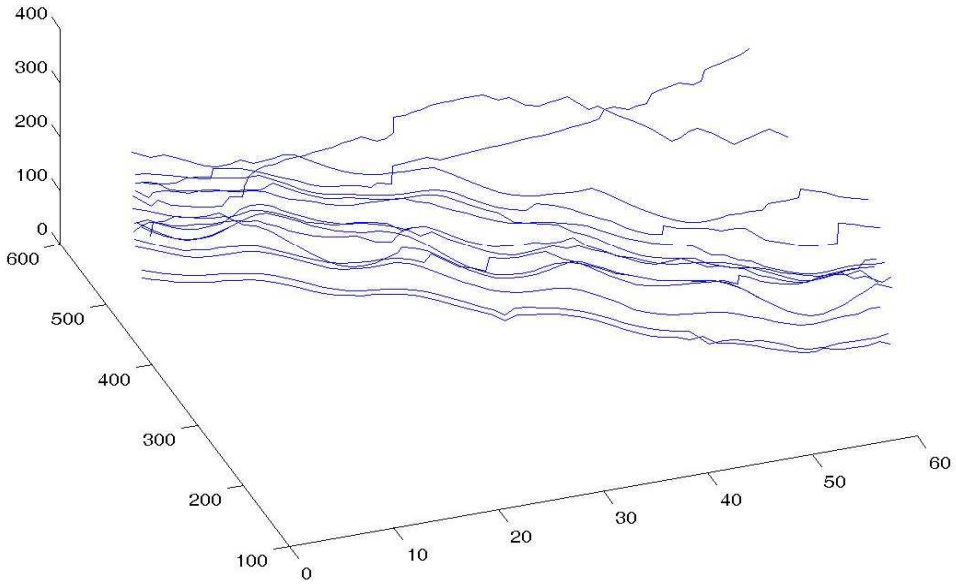


(v. g)



(v. h)

Y- Axis



X - Axis

Fig (vi)

Time t

Figure (vi) shows 3d plot of tracked points for running

The figures from (vii. a) – (vii. g) below shows the tracking pixels from first frame till the last frame, actor is walking from right to left. And the X and Y denote the axis that represents pixel position

fig (vii. a)

fig (vii. b)

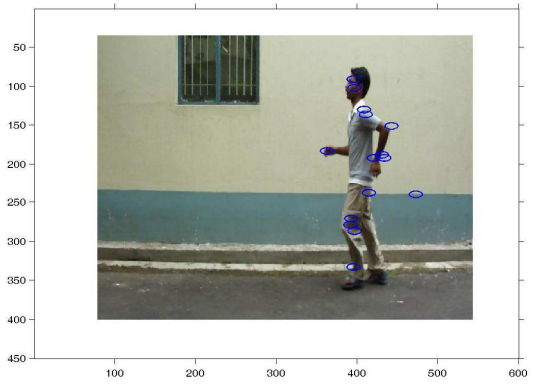


fig (vii. c)

fig (vii. d)

fig (vii. e)

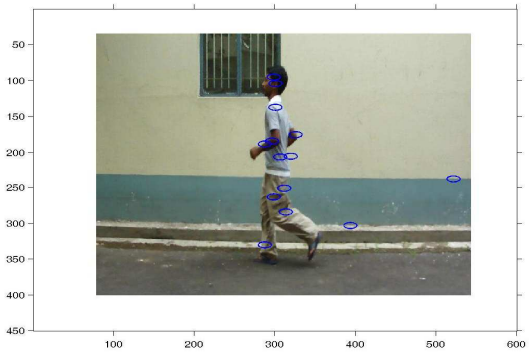


fig (vii. f)

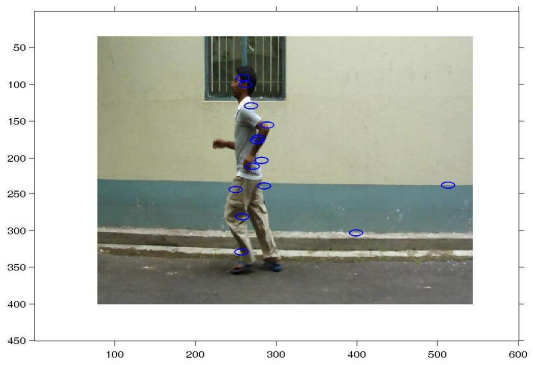


fig (vii. g)

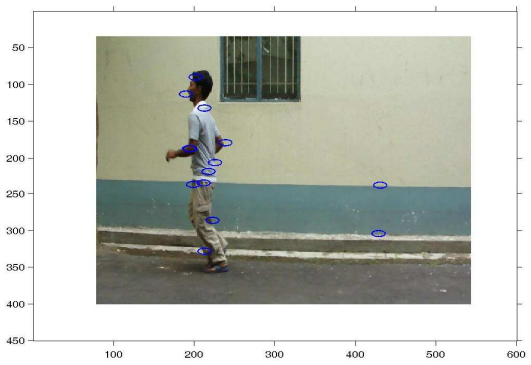


fig (vii. h)

dct(f)

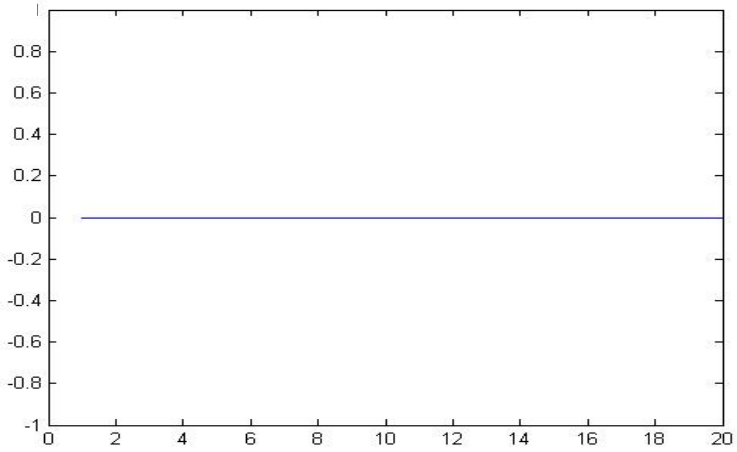


Fig (viii) Frequency f

The figure (viii) shows the direct cosine transform of the horizontal strip that does not have any pixel of interest

dct(f)

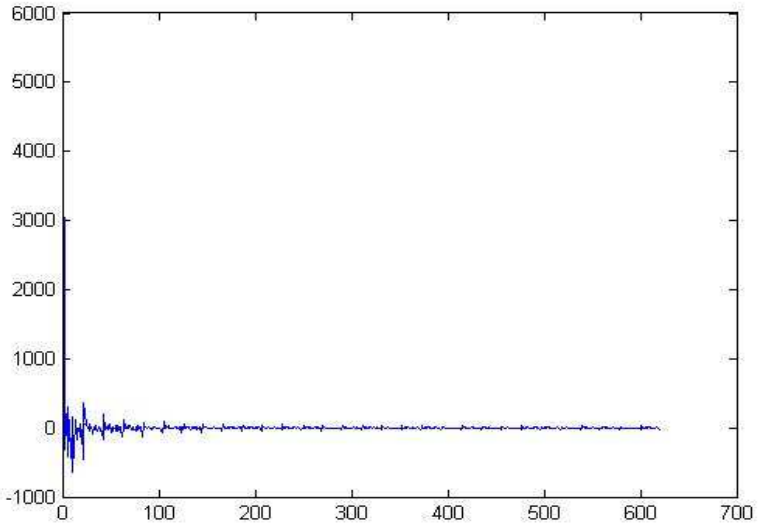


Fig (ix) Frequency f

The fig (ix) shows the direct cosine transform of the horizontal strip that contains at least one pixel of interest. Here the points that we get from the direct cosine transform of the signal that has low frequency are taken as feature points.

The above Experiment is run for 30 times using leave one out strategy. The test set is chosen randomly, the results are not quite impressive, shown in tabular format given below

	Class –1					Class-2			
1	2	2	2	2	2	2	2	1	1
2	2	2	2	1	1	1	1	2	1
3	2	1	2	1	1	2	2	2	2
4	2	1	1	1	2	1	1	2	1
5	1	2	1	1	2	2	2	1	1
6	1	1	2	1	1	1	2	1	1
7	2	1	2	1	1	2	2	2	2
8	1	1	1	2	1	1	1	1	1
9	2	1	1	1	1	1	1	1	2
10	2	2	1	2	2	1	1	2	2
11	1	1	2	1	1	1	2	1	1
12	2	2	2	2	2	1	1	2	1
13	2	2	2	2	2	2	2	1	1
14	2	1	2	2	2	1	1	1	1
15	1	2	2	2	2	1	1	1	2
16	2	1	1	1	1	1	1	1	1
17	1	1	2	2	2	2	2	2	2
18	1	1	2	1	1	2	2	2	2
19	2	1	2	2	2	1	2	2	2
20	1	1	2	2	1	1	2	1	1
21	2	2	1	2	1	1	1	1	1
22	1	1	2	2	1	1	1	2	2
23	1	1	1	1	1	1	1	2	1
24	1	2	2	2	2	2	1	1	1
25	1	1	1	2	1	2	2	1	2
26	2	2	1	1	2	1	1	1	1
27	1	1	2	1	1	1	1	2	1
28	1	1	1	1	1	1	2	1	2
29	1	2	2	1	2	2	2	2	1
30	2	2	1	2	1	2	2	2	2
# of points misclassified	15	12	18	15	13	19	15	16	18
Ratio	50%	40%	60%	50%	43%	63%	50%	53%	60%

Fig(x)

Fig (x) shows the misclassified points and the misclassification ratio

Chapter 5

5.1 Conclusion

The method we used here will track the object correctly only when the shape does not change. if there is any change in the shape then we may miss the tracking of an object.

Our project has the following disadvantages

User should manually select the points or region of interest. As we have used only correlation coefficient for the best matching region of interest some of the points may be missed when the regions are rotated or change its shape

Future Work

The method can be improved in future by using the fast and good block matching algorithms. And can also use more than one algorithm for goodness evaluation for the block matching criteria. If possible we try to extract many good features from the signal can also improve the performance of system.

Bibliography

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding, CVIU*, 73(3):428–440, March 1999**.**

[2] Moshe Blank, Lena Gorelick, Eli Schechtman, Michal Irani, and Ronen Basri.
Actions as space time-shapes. In *IEEE International Conference on Computer Vision, ICCV*, volume 2, pages 1395–1402, Oct 2005.

[3] Aaron F. Bobick and JamesW. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Analysis Machine Intelligence, PAMI*, 23(3):257–267,Mar 2001.

[4] D. M. Gavrila. The visual analysis of human movements: A survey. *Computer Vision and Image Understanding, CVIU*, 73(1):82–98, January 1999.

[5] Ju Han and Bin Bhanu. Individual recognition using gait energy image. *IEEE Trans. Pattern Analysis Machine Intelligence, PAMI*, 28(2):316–322, Feb 2006.

[6] Vision-Based Human Tracking and Activity Recognition Robert Bodor, Bennett Jackson, Nikolaos Papanikolopoulos, AIRVL, Dept. of Computer Science and Engineering, University of Minnesota

[7] D. Mahajan, N. Kwatra, S. Jain and P. Kalra, "A Framework for Activity Recognition and Detection of Unusual Activities", *ICVGIP*, 2004

[8] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden Markov models ", *ICCV*, 2003.

[9] B. Laxton, J. Lim and D. Creigman, "Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video", *CVPR*, 2007.

[10] C. Pinhanez, "Representation and Recognition of Action in Interactive Spaces", *Ph.D. Thesis*, 1999.

[11] Y. Shi, Y. Huang, D. Minnen, A. Bobick and I. Essa, "Propagation Networks for Recognizing Partially Ordered Sequential Actions ", *CVPR*, 2004

[12] R. Filipovych and E. Ribeiro, "Recognizing Primitive Interactions by Exploring Actor-Object States", *CVPR*, 2008

[13] C. Rao, A. Yilmaz and M. Shah, "View-Invariant Representation and Recognition of Actions", *IJCV*, Vol.50, Issue.2, 2002.

[14] O. Boiman and M. Irani, "Detecting Irregularities in Images and in Video", *ICCV*, 2005.

[15] S-FWong, T-K. Kim and R. Cipolla, “Learning Motion Categories using both Semantic and Structural Information”, *CVPR*, 2007

[16] C.Stauffer andW.E.L.Grimson. Adaptive background mixture models for realtime tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, volume 2, page 252, June 1999.

[17] R Cucchiara, C Grana, M Piccardi, A Prati, and S Pirrotti. Improving shadow suppression in moving object detection with hsv color information. In *IEEE Conference on Intelligent Transportation Systems*, pages 334–339, Aug 2001.

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled ***"Human Activity Recognition System"*** is an authentic record of the dissertation carried out by AravindKumar Yadav .A at Indian Statistical Institute Kolkata, under my supervision and guidance. The work fulfils the requirement for the award of the M-tech degree in Computer Science.

Dated:

.....
(Dr Dipti Prasad Mukherjee)
Supervisor

.....
Countersigned
External Examiner

Acknowledgement

It has been a great honour and rewarding experience to work under the auspices of a guide as Prof Dipti Prasad Mukherjee, ECSU. It has been Prof Dipti Prasad Mukherjee's effort and encouragement that has borne fruits in the successful completion of this project. I consider myself extremely fortunate to have a chance to work under his supervision. No amount of thanks can repay his contribution to this work.

I also thank whole heartedly Prof B. Chanda all the faculty members of the ECSU for the invaluable knowledge they have imparted to me in the most exciting and enjoyable way. I would like to thank Snehashis, Sithanshu, Sujoy and the other members of ECSU at ISI Kolkata for their timely help and suggestions for my project. I also extend my thanks to the staff of the department and rest of the institute for being supportive during my entire stay and maintaining excellent working facility.

I thank all my classmates and friends Ravikishore, Chiru, Murthy, Vikram, Avathar for sparing time and being the subjects in the video shooting required for my project. I thank my wonderful friends at ISI Kolktata campus for standing by me in the difficult situations during my stay.

I am fortunate enough to be associated with an institute of international repute that has provided me an excellent infrastructure and environment in helping me to complete my project. I would like to thank my parents for supporting me through every inch of my career, through all the successes and failures. It was their blessings which always gave me courage to face all the challenges and made my path easier.

AravindKumar Yadav .A
Indian Statistical Institute, Kolkata

Contents