# HUMAN WALKING MOTION DETECTION AND CLASSIFICATION OF ACTIONS FROM VIDEO SEQUENCES

Dr. J. Narendra Babu[1], Dr. M Kezia Joseph[2], Dr N. Rajesha[3], Dr. B. Jayachandran[4], Dr.Nikhil Raj[5]
[1,2,3,4,5]Professor, Dept. of ECE,MRCE, Hyderabad

**Abstract:**

**Automatic recognition and analysis of human actions in videos helps in improving tasks relating to many manually demanding tasks relating to video surveillance, retrieval of videos from large databases, pedestrian traffic monitoring, and many others. This paper presents silhouette based human motion detection and classification of actions performed by humans in video sequences. The typical goal of this paper is to demonstrate understanding of human walking behavior which is a high-level task relying on several, lower-level tasks such as human detection, classification, tracking and pose estimation for walking, running, jumping, falling etc. Experiments were done on different test video sequences to demonstrate the efficiency in detecting silhouettes and classifying different actions using SVM Classifier.**

**Keywords- moving human detection, feature set, HOG, SVM, action recognition**

## I. INTRODUCTION

The topic of analyzing actions of humans found in video sequences in general has several applications in fields ranging from human motion analysis in clinical studies and sports medicine, to animation for both cinema and video games, and human-computer interaction. The current generation of video surveillance systems is geared towards recognizing actions of interest automatically from video streams such as tampering with infrastructure, drawing graffiti, looking inside a car, etc. To this aim, computer algorithms can be employed to learn to recognize such actions simply from presentation of example videos. However, the learning time can prove excruciatingly high, up to the order of days or weeks of computing time depending on the size of the dataset. However, it is a common opinion that many open issues still affect the accuracy of action recognition. As a main challenge, the instances of the same action by various people are significantly different; moreover, every individual performs each action in a different manner over various instances, both in space and time. This can be formulated as a problem of high, intrinsic within-class variability. Adding to the challenge; the number of samples available for training is typically limited compared to the parameters, preventing a "brute force" training approach. Human detection and tracking are challenging tasks as human body is highly articulated and people tend to wear complex clothing texture.

As an initial study in this paper, we have taken up the human walking motion analysis to capture various actions that a human could perform while walking such as running, jumping, falling etc. To analyze human walking motion which is a high-level task we have proposed a pipeline of tasks that rely on several, lower-level tasks such as human detection, classification, tracking and pose estimation. This paper presents a method useful for the development of automated video surveillance systems. The typical goal of automatic action recognition is based on accurate classifications of actions in given image sequences as one of several classes of predefined actions.

## II. RELATED WORKS

Many computer vision applications require the

segmentation of foreground from background as a prelude to further processing tasks. Although difficult in the general case, the task can be greatly simplified if the object or objects of interest in the foreground move across a static background. Such situations arise or can be engineered in a wide variety of applications, including security videos, video-based tracking and motion capture, sports ergonomics, and human-computer interactions via inexpensive workstation-mounted cameras. All of these applications rely on or would benefit from high-quality foreground segmentation. Unfortunately, existing methods sometimes prove unreliable and error-prone. Furthermore, the results can vary greatly between successive frames of a video.

Graph-cuts is a popular technique used in object segmentation. It is able to overcome difficult scene conditions by looking at neighborhood similarities, grouping pixels together that would otherwise have been wrongly classified. In Grow cut algorithms the user defines the object of interest by drawing some rough strokes inside it with an object brush and some other strokes outside it using a background brush. The algorithm considers these as seed points and proceeds. Thus in graph cut algorithms multiple strokes acts as seeds while in region growing algorithm only one seed point is handled at a time.

Region growing is a simple region-based image segmentation method. It is also classified as a pixel-based image segmentation method since it involves the selection of initial seed points.

This approach to segmentation examines neighboring pixels of initial seed points and determines whether the pixel neighbors should be added to the region. The process is iterated on, in the same manner as general data clustering algorithms.

One of the other techniques is the MB-HOT feature. They are 3 pixels combination in a 3×3 region. For each template, if the intensity value of P is greater than the other two, it is regarded that the 3×3 region meets this template. An 8 bins histogram can be calculated for detection window. Each bin corresponds to one template. The value of each bin is the number of 3×3 regions meeting corresponding template in detection window. The main contribution includes: a multi scale block histogram of template feature is proposed. Texture information and gradient information are encoded in this feature and it can reflect the relationship of three blocks [2].

SeonHeo,Hyung Il Koo,Hong Il Kim,NamIk Cho [3] proposed a method which is based on energy minimization framework. Although some methods built 3D graphs and solved a 3D energy minimization problem for video segmentation, here it formulates the video segmentation problem as a frame-by- frame image segmentation problem. It is because to satisfy real-time requirements on video call applications. But this only uses the detection of faces in video call application.

Another system of motion capture that includes both the model acquisition and the motion tracking is by using 3D voxel data. While the approach of working directly with the image data is common, the 3D voxel reconstructions of the human body shape at each frame as input to the model acquisition and tracking. The system does not perform the tracking directly on the image data, but on the 3D voxel reconstructions computed from the 2D foreground silhouettes. This approach removes all the computations related to the transition between the image planes and the 3D space from the tracking and model acquisition [4].

Thinning algorithm is an algorithm which works on indoor video sequences to construct 2D human body model. This algorithm works on straight poses acquired by single static camera without using markers on the human body. Thinning algorithm plays a vital role in finding the skeleton of human body. The thinning operation is performed by transforming the origin of the structuring element to each pixel in the image. Then it is compared with the corresponding image pixels. When the background and foreground pixels of the

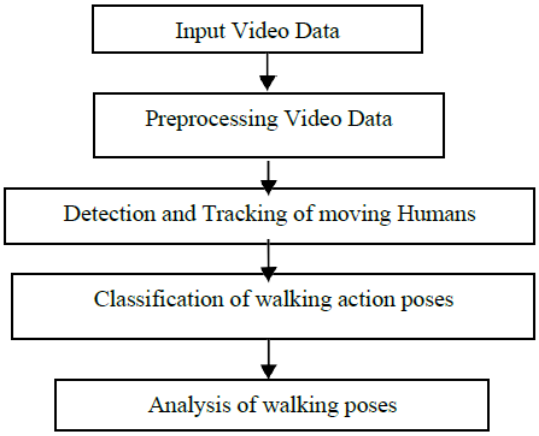
structuring element and images are matched, the origin of the structuring element is considered as background. Otherwise it is left unchanged. Here, the structuring element determines the use of the thinning operation. The structuring element determines the number of pixels added or deleted from the objects in an image [5].

There exist a number of algorithms to estimate the pose from images captured from a single camera. Segmentation of the image into different, possibly self-occluding, body parts and tracking them is an inherently a difficult problem. It is, therefore, recommended to use multiple cameras to deal with occlusion and kinematic singularities. In general, most computer vision algorithms target applications where only an approximate estimate of the pose is required. They also assume the human body model parameters are available. Usually, mathematical body models are used to deal with the large number of body segments and to guide the tracking and pose estimation processes [1].

Many different approaches for action recognition have been proposed over the past two decades. Any action recognition approaches requires the extraction of informative measurements from the video and their ensuing classification. In terms of classification approaches, two main approaches have been adopted: 1) recognizing the action directly in the time domain 2) recognizing the action by graphical models. The time domain approach has dynamic time warping (DTW) as its main representative, while the HMM is the reference generative approach for graphical models. [7] Other graphical models such as dynamic Bayesian networks and conditional random fields have also been used with significant degree of success. The approaches based on classification of histograms convert the time series of measurements, which varies in length for every action instance, into a histogram of values and then apply any conventional classifier such as the support vector machine or nearest-neighbors for the classification stage. Such histogram-based approaches have reported significant empirical accuracy.

## III. PROPOSED METHOD

The general framework proposed in our paper is as follows:




### A. Preprocessing Video Data

In this work initially, the video sequence used for the experimental purpose is to be converted to *. avi format and split into video frames of frame width 320 and frame height 240 i.e., (320x240)

### B. Detection and Tracking of moving Humans

To segment moving objects in video, motion detection is commonly used. In this work, the Mixture of Gaussian (MoG) technique that models each pixel as a Gaussian Mixture Model (GMM) will be used. This helps in foreground extraction along with the illumination differences and shadow removal. The sequential observations of data in frames are modeled based on two aspects.

1) Sequentiality (or proximity in time): Observations which are close in time are more likely to belong to the same state and, therefore, being drawn from the same GMM.

2) Proximity in feature space: each observation has a value, irrespectively of its occurrence in time. In general, those observations whose values are close to each other are more likely to be generated from the same state. Each observation is a multivariate measurement encoding the pose of the human in one video frame.

A bounding box will be placed over the moving object as the silhouettes are tracked in

the subsequent frames.

C. Classification of walking action poses

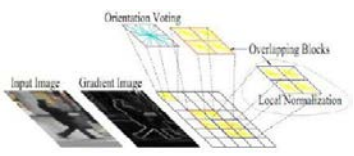
1) Feature set extraction using HOG

The method is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. The basic idea is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. In practice this is implemented by dividing the image window into small spatial regions ("cells"), for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combined histogram entries form the representation. For better invariance to illumination, shadowing, etc., it is also useful to contrast-normalize the local responses before using them. This can be done by accumulating a measure of local histogram "energy" over somewhat larger spatial regions ("blocks") and using the results to normalize all of the cells in the block. We will refer to the normalized descriptor blocks as Histogram of Oriented Gradient (HOG) descriptors. Tiling the detection window with a dense (in fact, overlapping) grid of HOG descriptors and using the combined feature vector in a conventional SVM based window classifier gives our human detection chain [9].

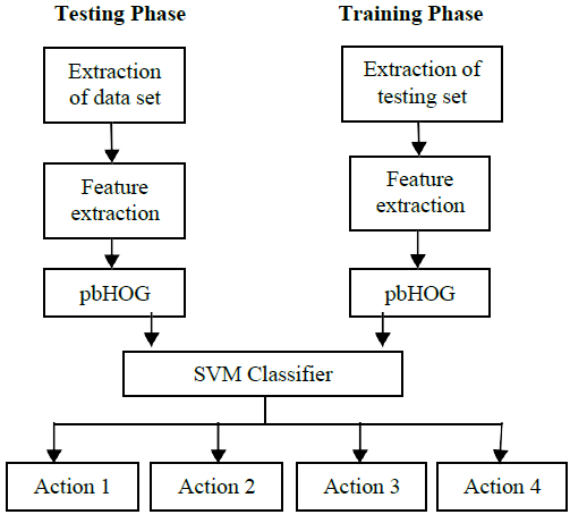
Figure 2: HOG feature extraction process of human walking motion in video frames.[8]

2) Classification using SVM

There are discriminative classifiers which try to find characteristic differences between positive and negative examples. SVM is a widely used technique. SVM computes a high-dimensional hyper plane to separate the different object categories. To compute the plane, the chosen image feature space or a kernel of this feature space is used. SVM performs a two-class classification in two stages, first is the training stage and second is the testing stage.




In this paper we propose to recognize actions by classification of histograms of the measurements. SVM is used for classification; it should be trained by sample data before classifications. The sample data is the HOG feature vectors of general human and background images. When the support vectors are trained sufficiently, the SVM classification can be used to recognize humans from static images or real-time video.




D. Analysis of walking poses:

The proposed analysis of walking poses can be performed using neural network for supervised classification. Action representation can be done by a self-organizing neural network training followed by fuzzy vector quantization. Action classification can be performed by a feed forward neural network which is trained for view-invariant action recognition. Multiple action classification results combination based on Bayesian learning, in the recognition phase, results to high action recognition accuracy.

The neural network generates a series of values for each of the action and then compares with the action of each frame.

Let us denote by {xi, ci}; i= 1; : : : ;N a set of N vectors xi $\epsilon$ D

followed by class labels ci$\epsilon${1; : : : ;C}.

We would like to employ them in order to train the network. Such a network consists of D input (equal to the dimensionality of xi), L hidden and C output (equal to the number of classes

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-12, 2017
DOI: 0.21276/ijcesr.2017.4.12.12

involved in the classification problem) neurons. The number of hidden layer neurons is usually selected to be much greater than the number of classes i.e., L C.

The network target vectors ti= [ti1;:::;tiC] T, each corresponding to a training vector xi, are set to tik= 1for vectors belonging to class k, i.e., when ci= k, and to tik= −1 otherwise. The network input weights Win ϵR D*Land the hidden layer bias values b ϵ RL are randomly assigned, while the network output weights Wout ϵ RL*C are analytically calculated. For a given activation function for the network hidden layer Φ(•) and by using a linear activation function for the network output layer, the output oi= [o1;:::; oc] T of the network corresponding to xi is calculated by

$$o_{ik} = \sum_{j=1}^{L} w_{kj}\, \Phi(\mathbf{v}_j, b_j, \mathbf{x}_i),\ k = 1, ..., C.$$

## IV. EXPERIMATAL RESULTS

The experiment of the human walking analysis in videos was executed in Matlab 2012. We use a preprocessed video having a single person and static camera for the action recognition.

A.Human detection and tracking snapshot

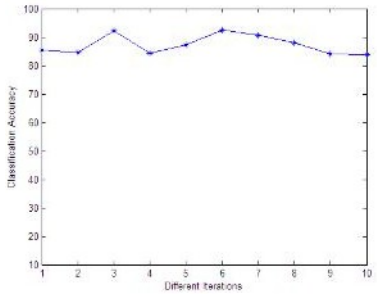



Figure 4: Tracking human during various activity a) Falling b) Sitting c) Jumping

B. Actions Classified:

For the SVM classification we used images of various actions about 50 in number for each action.

Training Phase

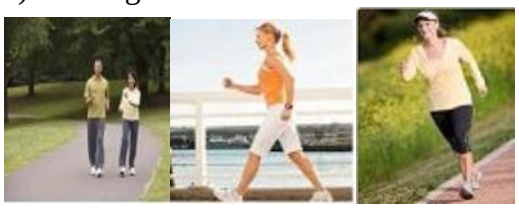
The data set for training the SVM consists of images of different actions. The feature extraction is done on the images. The pbHOG feature is then extracted for each of the image. The training instance matrix is formed from the feature values extracted and this is given to the SVM Classifier for training.

Testing Phase

The test data set is in the form of video which comprises of images for various actions. The video is first split into frames. The feature extraction is done on each frame. The pbHOG features are then extracted. The svm predictor then classifies the test data into different actions.

The various actions are classified as walking, running, sitting and dancing.

a)Walking



b) Running

c)Sitting

d) Dancing

Figure 5: Classification of actions using SVM Classifier a) Walking b) Running c) Sitting d) Dancing

Classification Results

The classification is iterated for a number of iterations and the accuracy for each of the iterations is plotted as shown in the figure. This

gives the efficiency of the method used.

Figure 6: Classification accuracy using SVM classifier

On running the test data set the amount of accuracy obtained for each of the class actions are plotted as a graph which gives the precision for the obtained result of classification in each category.

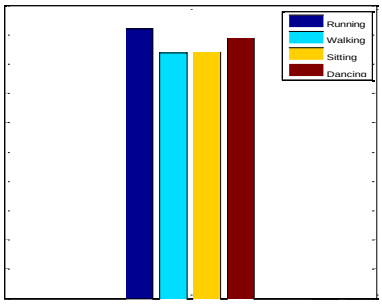



Figure 5: Classes of human action taken for experimentation

## V. CONCLUSION

In this paper, for human walking motion analysis the work presented is focused on human Motion detection and tracking is used to extract moving silhouettes later and actions areclassified using SVM Classifier. Actions are categorized into pre-defined groups using SVM classifier. Experimental results from real-time video are provided to show the effectiveness of the method. The experiments results prove the validity and effectiveness. In future work, this classification method will be used in intelligent surveillance field.

Analysis of walking motion is not demonstrated in this paper and is kept as future work. Neural networks will be used to classify the actions using a range of values that are generated for each of the action.

## REFERENCES

[1] Chen, Daniel and Denman, Simon and Fooke s, Clinton B. and Sridharan Sridha, “Accurate Silhouettes for Surveillance - Improved Motion Segmentation using Graph Cuts”, International Conference on Digital Image Computing : Techniques and Applications (DICTA 2010), 1-3 December 2010, Sydney, Australia.

[2] ShaopengTang,SatoshiGoto,”Accurate Human Detection by Appearance andMotion”,IEICE trans. Fundamentals/commun./electron./inf. & syst., vol. E85-a/b/c/d, no. 1, January 2002

[3] Ivana Mikic,Mohan Trivedi,Edward Hunter,Pamela Cosman,” Human Body Model Acquisition and Tracking Using Voxel Data”, International Journal of Computer Vision 53(3), 199–223, 2003

[4] SeonHeo,Hyung Il Koo,Hong Il Kim,NamIk Cho ,”Human Segmentation Algorithm for Real-time Video-call Applications”,NIPA- 2013

[5] K. Srinivasan, K. Porkumaran and G. Sainarayanan, “Development of 2D human body modeling using thinning algorithm”, ICTACT Journal On Image And Video Processing, November 2010 ISSUE02

[6] MunWai Lee and Ram Nevatia, “Body Part Detection for Human Pose Estimation and Tracking”, Institute for Robotics and Intelligent Systems University of Southern California-IEEE - Motion and Video Computing (WMVC'07)

[7] Zia Moghaddam and Massimo Piccardi, Senior Member, IEEE , “Training Initialization of Hidden Markov Models in Human Action Recognition “,IEEE Transactions On Automation Science And Engineering, Vol. 11, No. 2, April 2014

[8] Hou Beiping and Zhu Wen, “Fast Human Detection Using Motion Detection”, Journal of Computers, vol. 6, no. 8, August 2011

[9] Navneet Dalal and Bill Triggs,”Histograms of Oriented Gradients for HumanDetection”,in CVPR,2005, 886-893