Human Motion Detection and Tracking for Video Surveillance

Prithviraj Banerjee and Somnath Sengupta
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur, Kharagpur 721302, India
Email: banerjee.prithviraj@gmail.com, ssg@ece.iitkgp.ernet.in

***Abstract*— An Automated Video Surveillance system is presented in this paper. The system aims at tracking an object in motion and classifying it as a Human or Non-Human entity, which would help in subsequent human activity analysis. The system employs a novel combination of an Adaptive Background Modeling Algorithm (based on the Gaussian Mixture Model) and a Human Detection for Surveillance (HDS) System. The HDS system incorporates a Histogram of Oriented Gradients based human detector which is well known for its performance in detecting humans in still images. Detailed analysis is carried out on the performance of the system on various test videos.**

I. Introduction

Automated Video Surveillance addresses real-time observation of people and vehicles within a busy environment leading to a description of their actions and interactions [1]. Technical issues include moving object detection and tracking, object classification, human motion analysis, and activity understanding.

A. Motivation and present state of research

A standard surveillance system attempts to recognize the regions of interest in a video scene, i.e. the moving entities in the scene. The classification of the moving entity forms a critical part of the system, as the subsequent modules analyze the moving entity based on whether it is a vehicle, a human or a group of humans.

Significant amount of work has already been reported in the field of surveillance, with each system employing its own unique classification technique. The W^4 System by Haritaoglu et al [2] is an extremely successful system reported in the literature for detection and tracking of people in a video sequence. It segments out the foreground pixels using a statistical background model and describes them as blobs. The blobs are classified into one of the three predetermined classes (single person, people in a group and other objects) using static silhouette shape and dynamic periodicity analysis. Cutler et al [3] presented a surveillance system which tries to detect and analyze the periodic motion in a moving entity using Time Frequency Analysis. It tries to classify the object into humans, running animal, vehicles etc depending on the periodic motion present inside the object. For example, a walking person exhibits significant periodic motion in his hand and leg regions.

On the other hand, object classification in still images is quite a well developed field. Dalal et al [4] presented a novel system for recognizing humans in images using Histograms of Oriented Gradients. The algorithm is based on evaluating normalized local histograms of image gradient orientation in a dense grid. The method tries to characterize local object appearance and shape by a general idea of the distribution of local intensity gradients or edge directions. The system normalizes the image to a standard size before analyzing it. Hence the object should compose a significant portion of the image. To systematically search for a object in a image, Qiang Zhu et al. [5] used a Cascade of Histogram of Oriented Gradients combined with ADABOOST learning to search for humans in an image by taking search windows at various scales and locations in the image. To do so in a video sequence would be computationally quite expensive, as there would be around 15 to 30 images per second to search in, depending on the fps of the video.

Hence, a need was felt for adapting an already existing object recognition system for still images, such that it could be easily deployed for recognizing objects in a video sequence.

B. Proposed Video Surveillance System

This paper presents a novel combination of an Adaptive Background Modeling system and the Histogram of Oriented Gradient based human detection system for application in video surveillance. The surveillance system presented in this paper can detect and track moving objects in a video sequence, and is resilient against temporal illumination changes. The system also adapts itself to long lasting changes in the background over time. The moving entities are further classified into Human and Non-Human categories using the Human Detection for Surveillance (HDS) System. A brief overview of the system is given in Fig. 1.

The foreground is extracted from the video scene by learning a statistical model of the background, and subtracting it from the original frame. The background model learns only the stationary parts of the scene and ignores the moving foreground. The system uses the Gaussian Mixture Model for modeling the background adaptively. Hence the motion regions are identified in the frame, which constitute the regions of interest (ROI) for our HDS system. The ROI might consist of a human figure, an animal or even a vehicle. The Histogram of Oriented Gradients algorithm is applied on the ROI to detect which category of object is present in the ROI (Fig. 2). This makes it unnecessary for the HDS module to search for human

figures in the entire image.

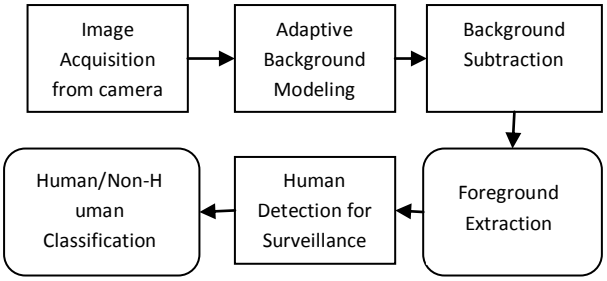

Figure 1: Surveillance System Overview

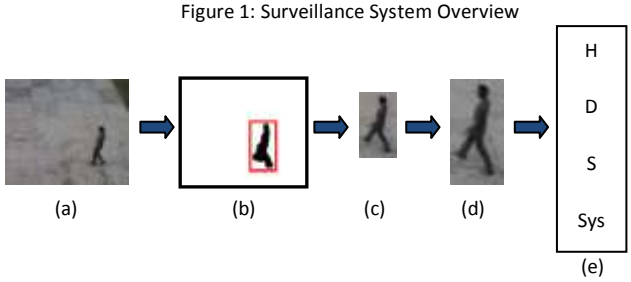

Figure 2: Application of HDS on Surveillance Video (a) Original Video Frame (b) Background Subtracted Frame with motion region detected (c) Extraction of image from motion information (d) Resizing image to 128X64 pixels (e) fed into HDS System for classification.

II. Adaptive Background Modelling for Motion Segmentation

A. Foreground extraction using Background Subtraction

A statistical background image of the video scene is obtained. This background image is subtracted from the current frame image and thresholded. The foreground regions of interest are extracted from the thresholded image after appropriate morphological operations. The algorithm flow for Static Background Subtraction is depicted in Fig 3.

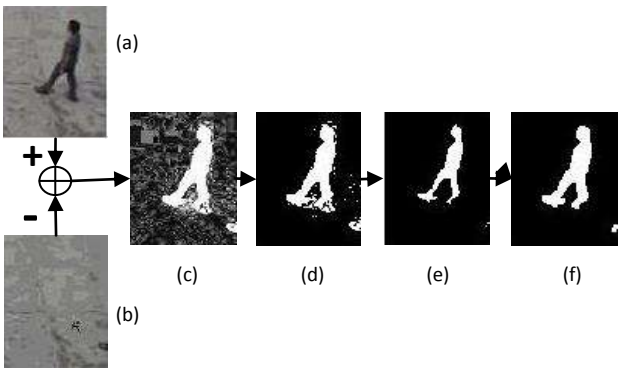

Figure 3: (a) Original Image (b) Estimated Background Image (c) Subtracted image (a)-(b). (d) Thresholding (e) Erosion (f) Dilation

The Static Background Subtraction system is not resilient to illumination changes or long lasting changes in the scene. Hence an Adaptive Background Modeling scheme should be adopted. In the following discussion, an implementation of the Gaussian Mixture model algorithm is presented, originally formulated by Stauffer et al [6], and subsequently modified by Harville et al [7].

B. Gaussian Mixture Model Overview

The following algorithm models each individual pixel X_t as a mixture of K-3D Gaussian distributions in the color space.

Pixel Value: $X_t = \begin{bmatrix} X_{r,t} \\ X_{g,t} \\ X_{b,t} \end{bmatrix}$ (1)

Probability: $P(X_t) = \sum_{i=1}^{K} w_{i,t} * \eta_{i,t}(X_t, \mu_{i,t}, \Sigma_{i,t})$ (2)

In each time step t, we try to match the new pixel observation $X_{i,t}$ with the K distributions in the mixture model. If we find a match, the weightings w_k, mean μ_k and variance $\sigma_{i,k}^2$ of the matched distribution $\eta_{i,t}$ are updated according to the equations described by Harville [7].

$$(X_t - \mu_{k,t})^2 > \beta\sigma_k^2 \quad (3)$$

$$\mu_t = (1-\alpha)\mu_{t-1} + \alpha X_t \quad (4)$$

$$\sigma_t^2 = (1-\alpha)\sigma_{t-1}^2 + \alpha(X_t - \mu_t)^T(X_t - \mu_t) \quad (5)$$

where α is taken as a learning constant. The distributions are sorted according to the values w_k/σ_k. The first B distributions are chosen from the sorting to represent the background according to the following criteria:

$$B = \arg min_b \left(\sum_{k=1}^{b} w_k > T\right) \quad (6)$$

The new pixel value is classified as a foreground pixel if no match is found amongst the B distributions. The least weighted distribution is replaced with the distribution corresponding to the new pixel value.

III. Human Detection for Surveillance (HDS) System

The algorithm description in the following section is based on the Navneet Dallal and Bill Triggs' *Histogram of Oriented Gradients (HOG) for Human Detection* [4]. Their implementation divides the image window into small spatial regions called cells. Each cell accumulates a local 1-D histogram of gradient directions or edge orientations of the pixel values in the cell. The histogram entries combine to give a unique representation for the image. Contrast normalization is also carried out across the cells to give better invariance to illumination changes.

A. Input Image Acquisition

The Adaptive Background modeling stage generates an approximate bounding box enclosing the motion object. The bounding box is resized to a dimension of 128X64 pixel size (Fig. 2c). The resizing makes the algorithm invariant to scale and size and hence people of varying heights and dimensions can be detected.

B. Gradient Computation

The Gradients in the image were computed using the simple 1-D mask [-1, 0, 1], in the vertical (V_g) and the horizontal (H_g) direction. The L1 norm is taken i.e. $\|Vg(x,y) + Hg(x,y)\|$ to compute the net gradient. Fig. 4 shows the resultant gradient computation.

C. Histogram Mapping and Binning into cells

The image is divided into multiple regions of fixed rectangular size called a cell. A histogram of gradient direction is computed for each cell, weighted by the gradient strength. The sign of the gradient is ignored as it only depends on the contrast difference between the foreground and the background in the image. In the HDS system, the range of $-90^o to\ 90^o$is divided into 9 equal Histogram bins (Table 1).

Table 1: Histogram Bin Mapping Table

Angular Range	Histogram Bin	Angular Range	Histogram Bin
$90^o\ to\ 80^o$	A	$0^o\ to -20^o$	F
$80^o\ to\ 60^o$	B	$-20^o\ to -40^o$	G
$60^o\ to\ 40^o$	C	$-40^o\ to -60^o$	H
$40^o\ to\ 20^o$	D	$-60^o\ to -80^o$	I
$20^o\ to\ 0^o$	E	$-80^o\ to -90^o$	A

Fig. 5(b, c) shows the resulting orientation computation for each pixel, with a darker shaded line representing a stronger gradient. In the HDS system, a cell size of 8X8 pixels is taken resulting in 8 cells width wise and 16 cells height wise. Fig 5(a) shows the net cell structure. Histograms of Gradients are computed over each cell, generating a 9 component feature vector.

Figure 4: (a) Original Image with Bounding Box and Resized Image. (b) Resized Image (c) Horizontal Gradient (d) Vertical Gradient (e) L1 Normalized Gradient

Original Image is taken from the INRIA dataset available at http://lear.inrialpes.fr/data

D. Overall Block Definition and Feature Vector Extraction

Local contrast normalization of the gradient values is desired to reduce the effect of illumination variations in the image. Normalization is achieved by spatially grouping the histogram cells (Fig. 5) into overlapping blocks. A single block consists of 2X2 cells and any two neighboring blocks have 2 cells in common due to block overlap. Each block consists of 4 Histogram of Gradient vectors, one from each of the four cells; resulting in a 36 Dimension feature vector. This feature vector is normalized using L1 Normalization scheme $[L1\colon v \rightarrow \frac{v}{\|v\|+\epsilon}]$. A 128X64 pixel image would consist of 7 blocks width wise and 15 blocks height wise, generating a 3780 Dimension feature vector.

Figure 5(d and e) shows the overall histogram of two blocks in the sample image. The left shoulder generates a histogram dominating towards the bins representing $60^o\ to\ 40^o$. Similarly the block over the right shoulder gives a histogram dominant in the range $-20^o\ to -60^o$. The histograms clearly capture the edge or gradient structure that is very characteristic of local shape.

E. Light Support Vector Machine – Training and Testing

The classification system in HDS employs a Support Vector Machine. HDS feature vectors are large in size (3780-Dimension long), and hence light Support Vector Machine (SVMLight [8]) were used to train the feature vectors. An implementation of the SVMLight from the MATLAB Arsenal library [9] was used in the HDS system. The SVM classifier decides whether the region of interest contains a human figure or not.

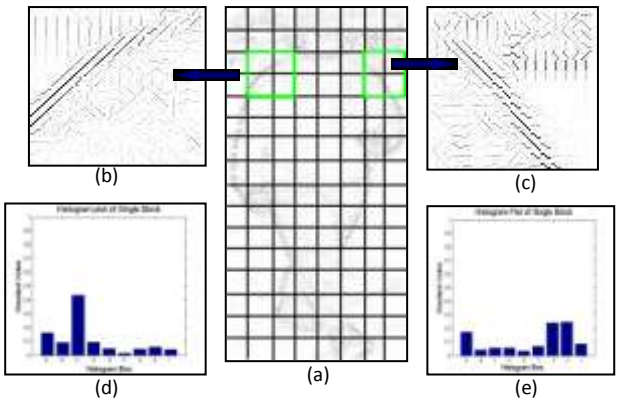

Figure 5: (a) Division of Gradient Image into Cells and Blocks. (b, c) Block consisting of 4 cells. (d, e) Overall Histogram from a Block of four cells. (Note: Here the histogram has been taken over all the 4 cells for representation purpose only. In the implementation, histograms are taken over individual cells, normalized and then concatenated together)

F. Incorporation of HDS system into Video Surveillance

The HDS system was initially trained on a few sample surveillance videos containing both positive (human) and negative (nonhuman like cars, animals etc.) moving objects. It was next tested on some testing videos. The HDS system makes a decision at each frame for a particular object being tracked. After a certain fixed number of frames, say 100 frames, have been analyzed for a tracked object, the HDS system computes the majority decision taken on the 100 frames and declares it to be a human or nonhuman entity.

IV. Results

A number of experiments were carried out for Adaptive Background Modeling using the Gaussian Mixture Models and the HMD system.

A. Experimental Setup

In the experimental setup, five Gaussian mixture models were considered for the Adaptive Background modeling. The learning constant α was taken to be 0.01. All sample videos were taken from Casio EXZ-750 Digital Camera at approximately 30 fps. The means of all the Gaussian distributions were initially set to zero and the variances were set

to 0.0016.

B. Test Results

The INRIA-Person image database was used for training and testing the Support Vector Machine. The database is available for download at http://lear.inrialpes.fr/data. A total of 615 positive trainings samples were taken from the INRIA dataset and 1200 randomly sampled regions (not consisting of humans) were taken as the negative training set. The trained SVMLight was tested on the test images provided in the INRIA-dataset and achieved a recognition rate exceeding 90 percent. 288 positive samples and 500 random negatives samples from the INRIA dataset are used for testing. The results are given in Table 2.

The HDS system was also tested on a number of videos of varying content. It was initially tested on a simple video consisting of a human figure walking across the frame. The system was able correctly detect the motion object and classify it as human with an accuracy measure of 85.7% (Table 2). The system was also tested on whether it could distinguish a human person from an animal in a video sequence. An example of an animal being tracked by the HDS system is shown in Fig. 6. The gradient image clearly shows a marked difference between the gradient arrangement of a dog and a human being (Fig. 4 and 5). This information is what enables the HDS system to distinguish between the two classes.

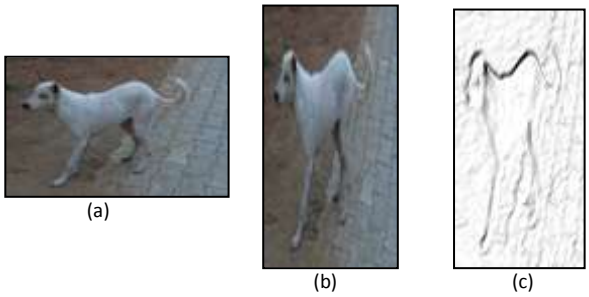

Figure 6: (a) Region of interest (b) Resized Image (c) Gradient Image

The support vector machine is trained as a binary classifier, and hence at a time can identify only a single category of objects. If the SVM is trained for humans, it will be able to classify the moving objects as humans or non-humans. A separate SVM is required to recognize the nonhuman as a dog or some other object. An anomaly occurs in the system if the motion segmentation module is not able to cleanly separate the bounding boxes of the human and the animal. This occurs, if both the human and the animal are very close to each other, with overlapping motion regions. In such a case, the Adaptive Background Subtraction module produces a single combined bounding box encompassing both the entities. When the HDS system trained for humans, is presented with such a combined bounding box, it will detect the presence of the human successfully and will ignore the presence of the animal. A second HDS system (trained for animals) will be required to recognize the presence of the animal. Also, if the motion bounding box encompasses more than one human (Figure 7), the HDS system will not be able to recognize each human figure separately. The HDS system only indicates the presence of a human figure in the region of interest, but will not be able to identify the quantity of such figures.

One of the key advantages that emerges from the test results is that the system does not require precise sleuth level motion segmentation, and can do with just an approximate bounding box around the moving entity. Also the system is able to run on low resolution videos, as it is predominantly based on overall gradient structures of the human figure, which would still be visible in low resolutions.

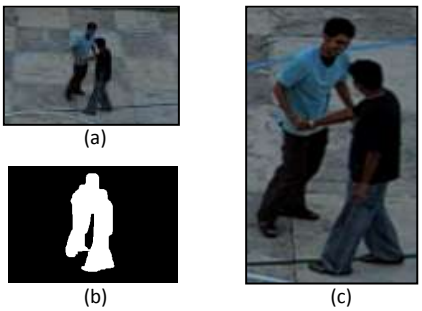

Figure 7: (a) Original Frame (b) Motion Segmentation (c) Resized Region of Interest

Table 2: Performance Results

Test Parameters	Values
Positive Test Images	
True Positives	222
False Negatives	66
Percentage True Positives	77%
Negative Test Images	
True Negatives	432
False Positives	68
Percentage True Negatives	86.4%
Surveillance Video	
Total Frames	540
Correctly Classified Frames	463
Accuracy Measure	85.7%

VI. Conclusion

A novel combination of Adaptive Background Modeling and Histogram of Oriented Gradients was presented for tracking and detecting human motion in a surveillance video sequence. The Human Detection for Surveillance (HDS) system was formulated, which could successfully classify a given image to be as Human or Non-Human in nature. The system used a Histogram of Oriented Gradients based approach for generation of feature vectors. The feature vectors were classified into the appropriate categories using a Support Vector Machine. The integrated system was successfully tested on a number of sample videos containing various moving entities. Analyses were made on its performance and certain key issues were identified to be kept in consideration while implementing the system.

REFERENCES

[1] Robert T. Collins, Alan J. Lipton, and Takeo Kanade, "Introduction to the Special Edition on Video Surveillance", *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 22, No. 8, pp 745-757, August 2000.

[2] Ismail Haritaoglu, David Harwood and Larry S. Davis, "W^4: Real-Time Surveillance of people and their Activities", *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 22, No. 8, pp 809-830, August 2000.

[3] Ross Cutler and Larry S. Davis, "Robust Real-Time Periodic Motion Detection, Analysis, and Applications", *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 22, No. 8, pp 781-796, August 2000.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human. Detection", *Proceedings of the Conference on Computer Vision and Pattern Recognition, San Diego, California, USA*, pp. 886–893, 2005.

[5] Q. Zhu, S. Avidan, M-C Yeh, , K-W Cheng, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients", *Proceedings of the IEEE Computer Society Conference on Computer vision and Pattern Recognition*, ISSN: 1063-6919, Volume 2, pp. 1491-1498, June 2006.

[6] C. Stauffer and W. Grimson, "Adaptive background mixture models for real time tracking", In *Proc. CVPR*, pp 246-252, 1999.

[7] M. Harville, G. Gordon, and J. Woodfill, "Foreground Segmentation using adaptive mixture models in color and depth", In *Proceedings of the IEEE workshop on Detection and Recognition of Events in Video*, 2001.

[8] T. Joachims, "Making large-scale svm learning practical", *Advances in Kernel Methods - Support Vector Learning*, The MIT Press, Cambridge, MA, USA, 1999.

[9] Rong Yan, MATLAB Arsenal- A Matlab Package for Classification Algorithms. http://finalfantasyxi.inf.cs.cmu.edu/MATLABArsenal/MATLABArsenal.htm School of Medical Science, Carnegie Mellon University.