# Moving Human Detection in Video Using Dynamic Visual Attention Model

G. Sanjay, J. Amudha, and Julia Tressa Jose

**Abstract.** Visual Attention algorithms have been extensively used for object detection in images. However, the use of these algorithms for video analysis has been less explored. Many of the techniques proposed, though accurate and robust, still require a huge amount of time for processing large sized video data. Thus this paper introduces a fast and computationally inexpensive technique for detecting regions corresponding to moving humans in surveillance videos. It is based on the dynamic saliency model and is robust to noise and illumination variation. Results indicate successful extraction of moving human regions with minimum noise, and faster performance in comparison to other models. The model works best in sparsely crowded scenarios.

**Keywords:** Moving Region Extraction, Visual Attention, Video Surveillance.

## 1 Introduction

Surveillance cameras are inexpensive and everywhere these days. However, manually monitoring these surveillance videos is a tiresome task and requires undivided attention. The goal of an automated visual surveillance system is to develop intelligent visual surveillance which can obtain a description of what is happening in a monitored area automatically, with minimum support from an operator, and then take appropriate actions based on that interpretation [1]. Automatic visual surveillance in dynamic scenes, especially for monitoring human activity, is one of the most active research topics in computer vision and artificial intelligence.

G. Sanjay · J. Amudha · Julia Tressa Jose
Department of Computer Science and Engineering,
Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bangalore, India
e-mail: {sanjaynair1989,julia.jose.t}@gmail.com,
j_amudha@blr.amrita.edu



El-Sayed M. El-Alfy et al. (eds.), *Advances in Intelligent Informatics*,
Advances in Intelligent Systems and Computing 320, DOI: 10.1007/978-3-319-11218-3_12

Extracting regions corresponding to moving humans is one of the most crucial initial steps in visual surveillance. It involves two steps – detection of moving regions and then extracting regions corresponding to humans. A typical approach first models the complex background, and then subtracts the background from each input frame to obtain foreground objects. Although the existing methods have achieved good detection results, most of them are computationally expensive. This factor is of utmost importance and should be kept to the minimum when a system has to be deployed in real time.

In the past few years, Visual Attention algorithms [2,3,4] have been extensively used in image and video processing as they are regarded to be computationally cost effective. These algorithms identify salient regions as foreground, allowing unimportant background regions to be largely ignored. By doing this, they enable processing to concentrate on regions of interest analogous to our human visual system.

Visual Attention has been extensively researched in images but there are relatively few works on video processing. Many models for images can be extended to include video data, however, not all approaches are fast, especially when the video content to be analysed has a large number of frames. Moreover, motion plays a crucial role in video and is more salient compared to other features such as colour and intensity.

This paper proposes a simplified technique for detecting moving humans in a video footage using dynamic visual attention model. The model processes more than a few hundred frames in less than thirty seconds and is robust to environmental noise and illumination.

Section 2 presents a literature survey on the existing visual attention models for moving object detection in video. The proposed model is presented in section 3, followed by results and conclusion in sections 4 and 5 respectively.

## 2 Visual Attention Models for Object Detection in Video

There are relatively few works which focuses on object detection in video using visual attention models. Salient objects in a video are identified using either static or dynamic attention cues. In a static saliency based model [5], the object stands out from its neighbours based on colour, intensity, orientation etc. No motion is considered. On the other hand, dynamic saliency based models [9, 10] give prominence to objects whose motion is salient to its background. Some models [6, 8], deploy a combination of static and dynamic cues to identify salient objects.

Some models [5, 6], rely on keyframe extraction as the initial step for video processing and object identification. This method summarizes the video content producing only the frames where relevant activity is found. In [5], this process is performed using the histogram keyframe extraction technique. A saliency map is then generated for the key frame using static attention cues based on the extended Itti Koch model [2]. This map indicates regions of relevance in a frame. The model selects human entities in the map through aspect ratio analysis. The average

aspect ratio of all salient objects forms the threshold value $T_h$. Regions having aspect ratios that exceed $T_h$ are identified as most salient by the system. The model has some difficulty in distinguishing objects having aspect ratio similar to humans.

There are models [6], which extract key frames through a combination of static and dynamic visual attention models (VAM), and thereby produce dynamic and static saliency maps. A VAI (Visual attention index) curve is obtained with the help of these maps. The peaks of the curve become the key frame candidate. For motion extraction from the frame, hierarchical gradient based optical flow method is used [7]. Finding two attention maps for each of the video frames becomes computationally expensive.

In contrast to the two methods discussed above, [8] does not use the concept of keyframe extraction at all. Instead, a motion attention map is generated by taking the continuous symmetry difference of consecutive frames. In addition to it, a static feature attention map and a Karhunen-Loeve transform (KLT) distribution map is computed. The final spatiotemporal saliency map is calculated as the weighted sum of these three maps. Each frame requires three attention maps to be obtained and summed which becomes a complex process when the number of frames is large.

In [9], a background reference frame is first created by averaging a series of frames in an unchanging video sequence. Motion vectors corresponding to moving regions are then found by calculating the intensity difference between current and reference frame. A second stage then applies a region growing and matching technique to these motion vectors to obtain motion segmentation. The method relies on an accurate background reference frame which may not be feasible in many cases. In the Mancas Model [10], motion features are extracted by making use of Farneback's optical flow algorithm. However, most optical flow methods are regarded as computationally complex and sensitive to noise.

Though the approaches discussed so far are based on visual attention, they still require extensive computations and a lot of time for processing large number of frames. Moreover only [5] discusses a method to extract humans from a video footage.

Therefore the aim of this paper is to propose a fast, simple yet robust technique to detect moving objects from a video and then segment regions corresponding to humans. The first step uses a modified version of the motion attention map proposed in [8]. For segmenting humans, aspect ratio analysis of [5] is used.

## 3 Moving Human Detection Using Visual Attention

This paper uses a modified version of the motion attention map, generated by using the continuous symmetry difference method [8]. It considers only motion features; no static cues about colours, grey levels or orientations are included in the model. The block diagram of the proposed approach is depicted in Figure 1.

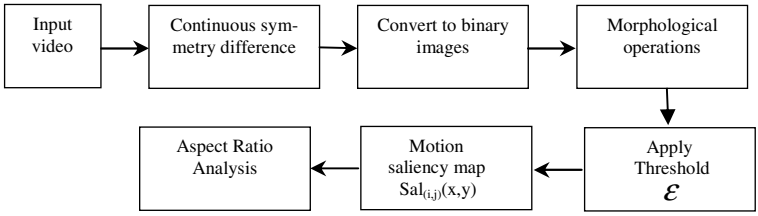



**Fig. 1** Block Diagram of Proposed method

The difference of adjacent images is calculated using the following formula:

$$dif_{(i,j)}(x,y) = |w^* I_i(x,y) - w^* I_j(x,y)| \quad (1)$$

where j = i+n, n, the frame difference, is chosen from 7-9 and w is a Gaussian filter function. Choosing such a high frame difference, does not lead to loss in information as the number of frames to be analyzed is more than a few hundred. This also speeds up the processing. The choice of $\sigma$ value of the Gaussian filter is also significant. A lower value reduces the overall noise; however there might be more holes in the moving region detected. A higher value reduces the holes but leads to increase in overall noise. The typical value of $\sigma$ ranges from 0.5-4.

The differenced image sequence is converted from RGB to binary. A few morphological operations (such as the imfill and imclose functions of MATLAB) are performed to fill the holes generated in the moving regions. To reduce output noise, values of difference less than a threshold ε are set to zero. For noise removal and to avoid missing moving object, ε is set to 1-2,

$$Sal_{(i,j)}(x,y) = \begin{cases} 1, \text{otherwise} \\ 0, \text{if } dif_{(i,j)}(x,y) < \varepsilon \end{cases} \quad (2)$$

The next important step is to separate moving regions corresponding to humans. Blobs having area greater than 100-150 are preselected to perform aspect ratio analysis. This value may vary depending on how far the camera is positioned from the actual scene. Aspect ratio of each detected blob can be calculated from its bounding box parameters.

$$Aspect\_Ratio(R_i) = \frac{\Delta y}{\Delta x} \quad (3)$$

where, Δy = Difference between two y extremes for the $i_{th}$ blob in a frame,

Δx = Difference between two x extremes for the $i_{th}$ blob in a frame.

This method is adapted from [5] where it is observed that human aspect ratio falls in the range 1 – 1.5. Blobs having lower aspect ratios are masked and thus eliminated.

In some cases an additional threshold $\theta$ (eqn. 2) can also be applied to the difference before conversion to binary, for further noise reduction. $\theta$ can be in the range 7-9. This alternate method is depicted in figure 2.

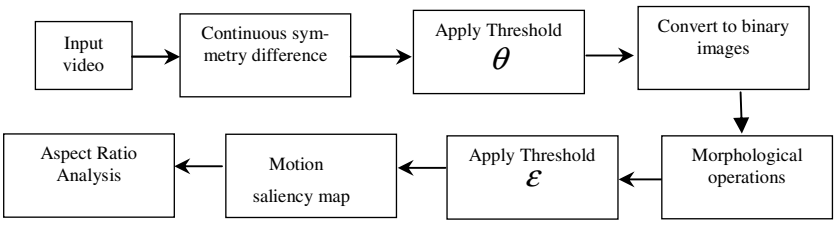



**Fig. 2** Alternate Block diagram (for improved noise reduction)

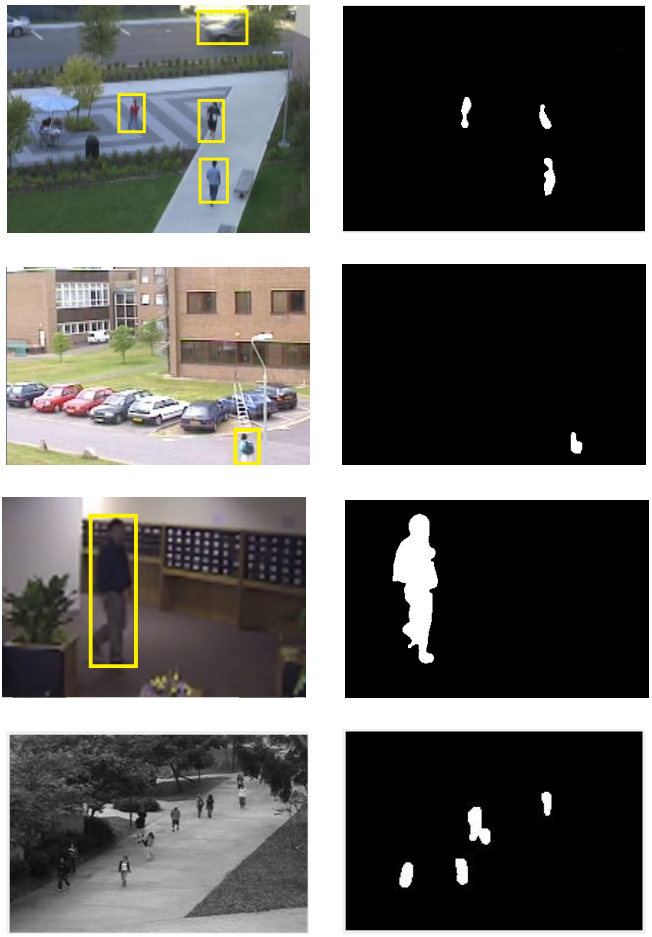
## 4 Results and Analysis

Four video sequences are chosen for evaluating the proposed method – video1, a video sequence of a public park [13]; video2, a part of PETS2001 dataset [12], and video3 with low lighting [10]; video4, a crowded scenario from the UCSD anomaly detection dataset [14]. All the experiments were conducted using MATLAB R2013a on an Intel Core i5-3210M CPU, running at 2.50 GHz.

Figure 3 shows the results of the proposed method. Regions corresponding to moving humans were successfully extracted from the first three videos in minimum time. The method provides good results even in dim lighting, as indicated by the second last row in fig. 3. The use of frame difference method contributes to the increased computational speed and better performance in low lighting. Setting a threshold value and filtering of detected blobs based on area and aspect ratio are the main factors which lead to decreased noise. However, as the crowd density increases, the model is not able to detect all the regions corresponding to moving humans, depicted in the last row of fig. 3. It works best in sparsely crowded scenarios.

Figure 4 shows a comparison of the execution times of the proposed method with [5], [6] and [10]. The proposed method took approx. 40 seconds for processing 700 frames of video1, whereas even a partial implementation of the other methods took more than a few minutes. This reduced execution time makes our method more suitable for real time applications.

Some models [5,6] identify relevant objects in a scenario by computing saliency map of the video frames. This approach was tested using two static attention models - Itti Koch[2]and Global rarity attention map[11]. The results in figure 5 indicate that the saliency maps contain other objects in the frames, in addition to moving objects. Extracting relevant moving objects from them proves to be a difficult task.



(a) (b)

**Fig. 3** Column (a)-Input: Moving Regions in video (indicated by yellow boxes); Column (b)-Output: Regions detected by proposed method. Each row indicates one of the four video datasets.

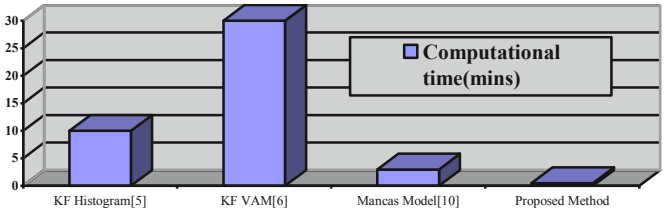
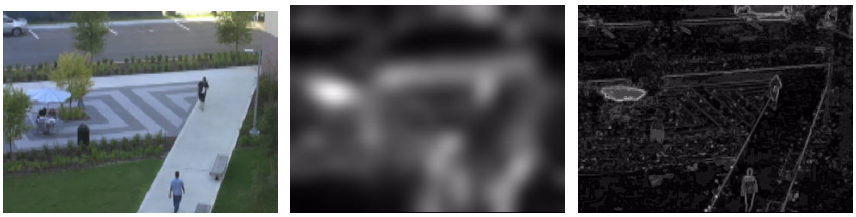



**Fig. 4** Comparing execution times of various models



(a) (b) (c)

**Fig. 5** (a) Original Frame, (b) Itti Koch map, (c) Global Rarity Map

## 5 Conclusion

A visual attention system for segmenting regions corresponding to moving humans is proposed in this paper. It is based on the dynamic saliency model where attention is mainly based on motion. The results obtained indicate that the model requires less time for computation compared to other models. It is quite robust to illumination and environmental noise. The model is more suited for simple and sparsely crowded scenarios. The values of the parameters: $\sigma$ of Gaussian filter and n, the frame difference may have to be adjusted for different scenarios. As a future enhancement the system can be implemented in hardware, for example in robotic vision applications and tested for computational efficiency.

## References

[1] Dick, A., Brooks, M.: Issues in Automated Visual Surveillance. In: Proceedings of International Conference on Digital Image Computing: Techniques and Application, pp. 195–204 (2003)

[2] Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis (1998)

[3] Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. LNCS (LNAI), vol. 3899. Springer, Heidelberg (2006)

[4] Amudha, J., Soman, K.P., Padmakar Reddy, S.: A Knowledge Driven Computational Visual Attention Model. International Journal of Computer Science Issues 8(3(1)) (2011)

[5] Radha, D., Amudha, J., Ramyasree, P., Ravindran, R., Shalini, S.: Detection of Unauthorized Human Entity in Surveillance Video. International Journal of Engineering and Technology 5(3) (2013)

[6] Amudha, J., Mathur, P.: Keyframe Identification using Visual Attention Model. In: International Conference on Recent Trends in Computer Science and Engineering, Chennai, pp. 55–55 (2012)

[7] Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on theory for warping. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)

[8] Guo, W., Xu, C., Ma, S., Xu, M.: Visual Attention Based Motion Object Detection and Trajectory Tracking. In: Qiu, G., Lam, K.M., Kiya, H., Xue, X.-Y., Kuo, C.-C.J., Lew, M.S. (eds.) PCM 2010, Part II. LNCS, vol. 6298, pp. 462–470. Springer, Heidelberg (2010)

[9] Zhang, S., Stentiford, F.: A saliency based object tracking method. In: International Workshop on Content-Based Multimedia Indexing, pp. 512–517 (2008)

[10] Riche, N., Mancas, M., Culibrk, D., Crnojevic, V., Gosselin, B., Dutoit, T.: Dynamic saliency models and human attention: a comparative study on videos. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part III. LNCS, vol. 7726, pp. 586–598. Springer, Heidelberg (2013)

[11] Mancas, M., Mancas-Thillou, C., Gosselin, B., Macq, B.: A Rarity-Based Visual Attention Map - Application to Texture Description. In: IEEE International Conference on Image Processing, pp. 445–448 (2006)

[12] Performance Evaluation and Tracking and Surveillance, PETS (2001), http://ftp.pets.rdg.ac.uk/PETS2001/DATASET1/

[13] Basharat, A., Gritai, A., Shah, M.: Learning object motion patterns for anomaly detection and improved object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)

[14] Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly Detection in Crowded Scenes. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA (2010)