# EXPERIENTIAL SAMPLING BASED FOREGROUND/BACKGROUND SEGMENTATION FOR VIDEO SURVEILLANCE

*Pradeep K. Atrey*[†]*, Vinay Kumar*[*]*, Anurag Kumar*[*] *and Mohan S. Kankanhalli*[†]

[†]School of Computing, National University of Singapore
[*]Indian Institute of Technology, Kharagpur, India

## ABSTRACT

Segmentation of foreground and background has been an important research problem arising out of many applications including video surveillance. A method commonly used for segmentation is "background subtraction" or thresholding the difference between the estimated background image and current image. Adaptive Gaussian mixture based background modelling has been proposed by many researchers for increasing the robustness against environmental changes. However, all these methods, being computationally intensive, need to be optimized for efficient and real-time performance especially at a higher image resolution. In this paper, we propose an improved foreground/background segmentation method which uses Experiential Sampling technique to restrict the computational efforts in the region of interest. We exploit the fact that the region of interest in general is present only in a small part of the image, therefore, the attention should only be focused in those regions. The proposed method shows a significant gain in processing speed at the expense of minor loss in accuracy. We provide experimental results and detailed analysis to show the utility of our method.

## 1. INTRODUCTION

Real-time processing of video has always been a problem in many applications including automatic video surveillance. In automatic video surveillance, one of the major steps in video based human-activity recognition is the foreground/background segmentation which takes substantial amount of computation time. In this paper, we focus on improving the computational efficiency of an existing foreground/background segmentation algorithm to meet the real-time requirements.

The core idea of our method is to use Experiential Sampling (ES) technique [1] to find the region of attention in each video frame and to restrict the processing to it. The ES technique utilizes the past experience to model the goal based contextual attention using which it finds the region where the computations need to be done. The goal, in our case, is to segment the foreground from background. The ES technique provides an efficient way to derive the attention samples from the media (sensor) samples. Once we have the attention samples, the processing is done only on the attention samples instead of the entire data. The ES technique has been shown useful in many applications including face detection and monologue detection in video [1]. We also exploit the temporal redundancy of the video to reduce the number of computations.

We have used adaptive Gaussian method to model the background as described by Stauffer et al. [2] and further improved by KaewTraKulPong et al. [3]. These methods do the computations on the whole image without taking into consideration of the regions of interest. It consumes a significant amount of time in doing unnecessary computations especially in non-busy environments where most of the frames captured by the camera has a clear background and should not be given much attention. The proposed method that integrates ES technique with the already proposed methods of background segmentation shows a significant improvement in the computational speed. This improvement in speed is achieved at the cost of minor loss in accuracy. This loss however is acceptable in light of the fact that any event lasts for sufficiently large number of video frames and the number of video frames in which the foreground is missed (in our method) is very less. And, of course, no surveillance task ends up at the segmentation of foreground only, rather it undergoes further analysis viz events detection, monitoring and tracking etc which relies on a series of video frames before concluding about an event. Hence even if the foreground in a few frames are missed by the detector, it does not affect the final objective appreciably. We have shown through experiments that the loss in accuracy is very small compared to the gain in computational speed.

Our contributions in this paper are summarized as follows. We have proposed an Experiential Sampling technique based foreground/background segmentation method which provides improved computational efficiency at the cost of negligible loss in accuracy.

## 2. RELATED WORK

Since we use both foreground/background segmentation as well as experiential sampling, we describe the related works in both of them. Background subtraction involves modelling a reference frame, subtracting the current frame, and then thresholding the result. This modelling though is simple but

is not robust enough to account for lightning changes in the scene. Koller et al. [4] used a Kalman filter to track the changes in background illumination for every pixel. In their method, only the most probable values of the background were included in the estimated background. Stauffer et al. [2] first came up with an adaptive modelling of background using Gaussian mixture which was robust enough for lightning changes and also to the new objects being removed or introduced into the scene which Koller's method lacked. But even this was not able to distinguish between foreground object and its shadow. Moreover, the method also suffered from slow learning. A solution to this was then proposed by KaewTraKulPong et al. [3] where they removed the likelihood term responsible for slow learning and used online Expectation Maximization algorithm for initialization, and then switching to the L-recent window update equations in order to give priority to the recent data and making the tracker adapt to the changes in the environment. But all these methods do the computations on the whole frame without taking into consideration the actual region of interest.

The concept of experiential sampling is inspired by the biological phenomenon of attention. In [1], a sampling based method has been used to represent the visual attention. The sampling based method provides a flexibility of representation, and can also be incorporated within a dynamical system which models the temporal continuity of visual attention. We have used similar method for modelling attention.

## 3. PROPOSED METHOD

The proposed method integrates experiential sampling technique [1] with the existing Adaptive Gaussian mixture method for foreground/background segmentation [2]. Experiential sampling is a process of selecting the most relevant information (i.e. Attention Samples) from the available data stream. The key idea is to concentrate in a direction which is most relevant and rewarding based on the context and past experiences. For example, if a person is walking in a corridor we need to focus our attention only in his vicinity instead of the whole image. But there is always a possibility of sudden change in context (for e.g. another person entering the scene) while the attention is already focused in a particular region. In order to account for such a change in the context, we must keep on refreshing the existing attention regularly. We do this rebuilding of attention profile at regular interval. We call this interval as "Attention Rebuild Window (ARW)". In our method, ARW is a significant parameter that bears a direct relation with the gain in processing speed and the incurred loss in accuracy. Hence, it needs to be properly tuned to achieve the optimization between the desired level of accuracy and processing time depending upon the scenario.

At any time $t$, the environment $e_t$ is modelled by -

$$e_t = \{S(t), A(t)\} \tag{1}$$

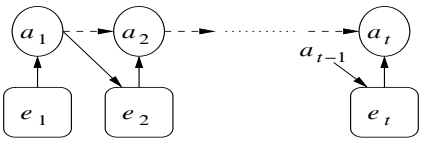


**Fig. 1**. Dynamic attention evolution model

where $S(t)$ denotes the *sensor samples* and $A(t)$ denotes the *attention samples*. The experiential sampling based approach that we have used, first considers the whole of image as sensor samples, and then builds the attention profile based on the region in which the foreground is detected. The attention samples dynamically evolve over the interval ARW and keeps on modelling the environment, and hence the attention samples in the incoming frames as shown in figure 1. To account for sudden change in context as discussed above, we rebuild the attention at regular intervals by throwing sensor samples in the whole of image.

We represent the sensor sample set as -

$$S(t) = \left\{s(t), \Pi^S(t)\right\} \tag{2}$$

where $s(t)$ is the set of $N_s$ number of pixel coordinates as, $s(t) = \{(x_1, y_1), (x_2, y_2) \dots (x_{N_s}, y_{N_s})\}$. $\Pi^S(t)$ is the associated weight or the importance of each sample given by, $\Pi^S(t) = \left\{\pi_1^S(t), \pi_2^S(t) \dots \pi_{N_s}^S(t)\right\}$, where each of $\pi_i^S(t)$ is obtained by fusing the spatial cues $C(t)$ available from the video data. In our case, we used the RGB color model and each of the values of RGB of a pixel served as cues. So, the cue set $C(t)$ is given by -

$$C(t) = \{c_r(t), c_g(t), c_b(t)\} \tag{3}$$

where,

$$\begin{aligned} c_r(t) &= \{(x_1, y_1, w_{r1}), (x_2, y_2, w_{r2}) \dots (x_{N_s}, y_{N_s}, w_{rN_s})\} \\ c_g(t) &= \{(x_1, y_1, w_{g1}), (x_2, y_2, w_{g2}) \dots (x_{N_s}, y_{N_s}, w_{gN_s})\} \\ c_b(t) &= \{(x_1, y_1, w_{b1}), (x_2, y_2, w_{b2}) \dots (x_{N_s}, y_{N_s}, w_{bN_s})\} \end{aligned}$$

and

$$w_{[r,g,b]i} = \begin{cases} 1 & \text{if it matches the Gaussian corresponding} \\ & \text{to its co-ordinate} \\ 0 & \text{otherwise} \end{cases}$$

Now, by employing the linear combination of sensor fusion strategy, we define the weights $\pi_i^s$ as -

$$\pi_i^s(t) = \sum_{j=r,g,b} \alpha_j . w_{ji} \tag{4}$$

where $\alpha_j$ is the importance of $j^{th}$ cue. The dynamically varying $N_a$ number of attention samples $A(t)$ are expressed by-

$$A(t) = \left\{a(t), \Pi^A(t)\right\} \tag{5}$$

where, $a(t) = \{(x_1, y_1), (x_2, y_2) \dots (x_{N_a}, y_{N_a})\}$ is the set of pixels within the bounding rectangle around the foreground

**Table 1**. The values set for different parameters

| Parameters | Range |
|---|---|
| ARW | 0-20 |
| $\alpha_i$ | 0.33 for each $i$, $1 \leq i \leq 3$ |
| $\xi$ | 1.5 |
| $\rho$ | 0 |

object. The associated weight $\Pi^A(t)$ of each attention sample is given by $\Pi^A(t) = \left\{\pi_1^A(t), \pi_2^A(t) \ldots \pi_{N_a}^A(t)\right\}$ and calculated as for sensor samples. The evolution model for the environment and attention is shown in figure 1. Initially, the environment $e_0$ consists of sensor samples $s_0$ (as the whole image) with no attention. Once a foreground is detected (say at time $t$) in the environment $e_t$, the bounding rectangle across foreground pixels becomes the attention samples $a_t$. The attention samples $a_t$ at time $t$ are used to compute the sensor samples $s_{t+1}$ at next time instant and hence construct the environment $e_{t+1}$. The function that maps previous attention to the new sensor samples is given by -

$$s_{t+1} = f(a_t), t > 0 \tag{6}$$

The $f$ is a linear function of the form, $f(x) = \xi.x + \rho$, where $\xi$ and $\rho$ are constants. $\xi$ denotes a scaling factor and $\rho$ denotes a displacement in the region of attention.

*Performance evaluation* is done by recording the computing time as well as the accuracy of the processing. The computing time and accuracy are compared with the existing foreground/background segmentation algorithm [2]. More specifically, for computing time, we measured the time (say $T_{normal}$) taken to process the total number of video frames using [2] (without ES), and also measured the time (say $T_{ES}$) taken by our method. The gain $G$ in processing time is computed as -

$$G = T_{ES}/T_{normal} \tag{7}$$

Similarly, for accuracy, let us say $N_{normal}$ and $N_{ES}$ are the number of video frames in which the foreground is detected using [2] and using our method, respectively; the loss $L$ in accuracy is given by -

$$L = (N_{normal} - N_{ES})/N_{normal} \tag{8}$$

## 4. EXPERIMENTAL RESULTS

The scenario in our experiment is a corridor with the camera at one end covering whole of the corridor. The objective is to perform foreground/background segmentation in real-time. The authors and other graduate students from our lab volunteered for performing walking, standing and running activities in the corridor. The 3000 video frames (in BMP format) of varying resolutions ($768 \times 576$, $384 \times 288$ and $160 \times 120$) were processed on an Intel P-IV 4 GHz processor. Various parameters used in the experiment are as shown in Table 1.

We compared our method with the existing methods (without ES) by analyzing the processing time and accuracy. The


**Fig. 2**. Blob detection results: The rectangle shows the number of attention samples in that frame.

processing time includes the time taken in foreground segmentation and the time consumed in connected component analysis and morphological operations (Erosion, Dilation). The accuracy is measured with respect to the existing method as mentioned in equation (8). The results obtained with and without experiential sampling are then compared both in terms of speed and accuracy.

The overall observations from the obtained results are-

1. Figure 2 shows the frames (190 to 250) of an event "A person walking in the corridor". Figure 3 illustrates a comparison between the two methods in terms of time consumed for processing these frames. We notice that, in the ES based method, the maximum time is taken in the processing of the frame where re-initialization of sensor samples is done, and subsequently the processing time is proportional to the number of attention samples in the frame. However, the processing time taken in the method without using ES is approximately constant and is much higher compared to our method.

2. As described earlier, the sensor samples are used to build the region of attention, figure 4 shows that the number of attention samples ($N_a$) closely follow the number of sensor samples ($N_s$) in the region between the peaks. The peaks correspond to the re-initialization of sensor samples at regular intervals (of ARW) to account for any contextual change. After the sensor samples are thrown on the whole of image, the evolved attention samples restrict the computations to the region of interest in the subsequent frames, thereby saving a lot of time and computation effort.

3. The experimental results show that there is a significant gain in overall speed for processing all the 3000 frames. In figure 5, we report the overall processing time of ES based method against the method without

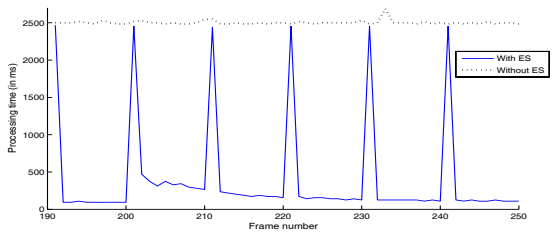


**Fig. 3**. Comparison of the two methods (Foreground/background segmentation with and without Experiential Sampling) in terms of processing time with ARW=10.

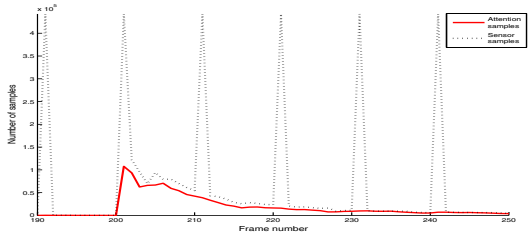


**Fig. 4**. Plot of sensor samples ($N_s$) and attention samples ($N_a$) versus the frame number with ARW=10.

ES by varying the ARW and the image resolutions. It can be clearly seen from the figure 5 that our ES based method is far superior in terms of average number of frames processed per second. For example, in case of $768 \times 576$ image resolution with ARW = 10, the average number of frames processed per second for ES based method are $\approx 3000/1200 = 2.5$, whereas it is $\approx 3000/4500 = 0.66$ for the method without ES.

4. Finally in figure 6, we present how the gain $G$ in processing time and loss $L$ in accuracy vary with ARW. The plot shows a substantial rise in gain in computation time at the cost of marginal loss in accuracy for

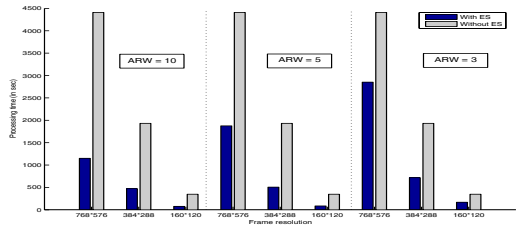


**Fig. 5**. Overall processing time comparison for varying frame resolutions and Attention Rebuild Window (ARW)

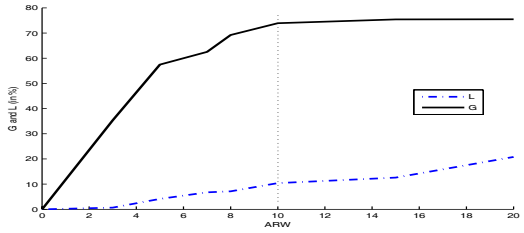


**Fig. 6**. Plot showing the percentage value of G (Gain in processing speed) and L(Loss in accuracy) as a function of ARW

smaller values of ARW. However, the gain curve tends to saturate for higher ARW while the error shows an increasing trend. This curve serves as the optimization curve for selecting an appropriate ARW given the maximum allowed inaccuracy (depending on the environment) and the speed requirements. For instance, a smaller ARW would yield better accuracy in crowed environments at the expense of gain in the speed.

## 5. CONCLUSIONS

The use of experiential sampling technique in the segmentation of foreground/background segmentation provides a substantial gain in processing speed (compared to the already proposed methods) to meet the real-time performance objectives at the cost of a minor loss in accuracy. Future work would be to test the method to handle multiple blobs in more crowded environment. It would also be interesting to study how Attention Rebuild Window (ARW) can adapt to the various environmental changes.

## 6. REFERENCES

[1] Kankanhalli M. S., Wang J., and Jain R., "Experiential sampling in multimedia systems," in *IEEE Transactions on Multimedia*, 2006, To appear. URL:http://www.comp.nus.edu.sg/∼mohan/ebs/.

[2] Stauffer C. and Grimson W. E. L., "Adaptive background mixture models for real-time tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999, vol. 2.

[3] KaewTraKulPong P. and Bowden R., "An improved adaptive background mixture model for real-time tracking with shadow detection," in *European Workshop on Advanced Video Based Surveillance Systems*, 2001.

[4] Koller D., Wever J., Huang T., Malik J., Ogasawara G., Rao B., and Russel S., "Towards robust automatic traffic scene analysis in real-time," in *IEEE International Conference on Decision and Control*, 1994, vol. 4.