Robust Motion Detection in Dynamic Indoor/Outdoor Scenes via Multi-layer Extraneous Motion Suppression

Wayne Chelliah Naidoo and Jules Raymond Tapamo

School of Computer Science
University of KwaZulu-Natal (Westville Campus)
Durban, South Africa
{naidoowc,tapamoj}@ukzn.ac.za

Abstract

Automated visual surveillance is poised to be a key technology in the fight against crime, particularly in monitoring security sensitive areas. However, its stability in realistic surveillance settings is highly dependable on robust detection of moving objects in a wide variety of environments. In this paper we present a background modelling approach that can handle complex scenes that are not completely static but contain small irrelevant motion. Our approach has the advantage that it is able to suppress these extraneous motions fast and efficiently, as well as adapt to changes in the scene which enables sensitive detection of active targets. In addition, it is capable of functioning in an array of environments both indoor and outdoor. We employ a coarse-to-fine motion suppression scheme. The initial coarse segmentation is constructed by a temporal median filter. A pixel process is then applied to learn the variability that each pixel exhibits over time. We represent this variability using a single adaptive Gaussian distribution. Using these per pixel distributions, we suppress any extraneous motion. The final level of suppression considers inter-pixel relations and filters artifacts left over from the previous step. As an example of the effectiveness of this approach, an inactive object warning module is presented. Our system runs in real-time. Initial experiments suggest that this approach handles sensitive detection with low false alarm rates.

Keywords: Motion detection, background subtraction, motion suppression, selective background update, inactive object warning.

1. Introduction

The emergence of new sensors and improved processing hardware in recent years has led to a strong demand for robust real-time automated visual surveillance systems [1]. The challenge is to reliably detect and track active objects in a wide variety of environments, both indoor and outdoor, which contain complex backgrounds with a minimal amount of user intervention and supervision. To this end, there has been a significant drive by researchers to develop algorithms that extend beyond indoor and controlled outdoor environments to more realistic dynamic environments typically encountered in real-world applications. The principal aim of motion detection is to isolate objects/regions of interest from the remainder of the observed scene whilst suppressing false positives caused primarily by background motion and lighting changes.

Background subtraction has been the foremost technique employed in segmenting moving objects in image sequences taken from a static camera. The pixels of interest are usually those that are the outliers to the model. The general motivation of partitioning an image into a set of $m$ non-overlapping regions $R \in \{r_1, \ldots, r_m\}$, is that the problem space is reduced from analyzing an entire complex scene to just analyzing its primitives. A simple and common motion detection method involves subtracting each new image $I_t(x, y)$ in a video sequence from a model of the background scene void of objects $B_t(x, y)$, and thresholding the resulting difference, highlighting foreground pixels (see Equation (1)). The most important component being the model of the background.

$$|B_t(x, y) - I_t(x, y)| > \theta$$

where $\theta$ is the threshold.

Haritaoglu et al. [2] in $W^4$, used a bimodal distribution model of background variation which was constructed during a training phase. The scene was modelled by representing each pixel by three values, its minimum and maximum intensity values and the maximum intensity difference between consecutive frames. Any pixel that deviated from its minimum or maximum by more than the maximum inter-frame change was considered foreground. The Wallflower project [3], applied a three component system for background maintenance. The pixel-level component performed Wiener filtering to predict pixel intensity given a recent history of values. The frame-level component applied temporal differencing to monitor sudden, global changes in illumination in order to reboot their model. However, a side effect of temporal differencing is an incomplete segmentation containing holes within the interior foreground pixels. Hence, a third region-level component was applied to fill in these homogeneous regions of foreground objects. Wren et al. [4], were the first to employ a single Gaussian to model the colour distribution for each pixel. Friedman et al. [5], extended this by explicitly classifying the pixel values into three separate distributions. However, these distributions where predetermined corresponding to the colours of the road, shadows and vehicles. In a general surveillance system, pixels may present a single background colour or multiple background colours resulting from repetitive motions, shadows or reflectance. Hence, any distribution should be automatically learned and maintained. Stauffer et al. [6], developed the mixture of Gaussians (MoG) approach. They modelled the recent history of each pixel with an approximation to a mixture of $K$ Gaussian distributions (typically, $K \in [3, 5]$). With the assumption that each new pixel value will be represented by one of the major components of the mixture model. Background subtraction was performed by marking any pixel that was more than $2.5\sigma$ standard deviations away from any of the weighted $B (B \subseteq K)$ distributions as a foreground.
pixel. Classification was based on most supporting evidence $\omega$ and least variance $\sigma$, since $\bar{x}$ increases both as a distribution gains more evidence and as the variance decreases. Many enhanced variants of the MoG have been proposed since. Some of them integrated colour, depth [7] and local features [8] into the Gaussians. Elgammal et al. [9] employed a non-parametric background model by estimating the probability of observing pixel intensity values based on a sample of intensity values for each pixel, using kernels, to replace the Gaussians.

The previously mentioned approaches are based on adaptive statistical modelling of pixel intensities. Although pixel intensity is not invariant to lighting changes, adaptive models make it possible to incorporate gradual changes in illumination. However, a sudden change in illumination presents a challenge to such models. The second major difficulty in modelling outdoor environments is that they often contain small irrelevant motion, that arise out of natural scene variations. These variations are often determined by the prevailing weather conditions such as swaying tree branches.

In this paper, we present an approach that can tolerate motion clutter arising out of natural scene variations as well as gradual illumination changes. The proposed framework can be applied to reliably detect objects of interest in general indoor/outdoor scenes. To illustrate the usefulness of this approach to real-time surveillance applications, an inactive object warning module is presented. By performing trajectory analysis, the system detects inactive objects. These may take the form of unattended or lost luggage, objects that were removed (e.g. a ghost region caused by a car leaving its parking bay), or changes to the scene (e.g. theft or vandalism).

The rest of the paper is organized as follows. In Section 2, the framework is introduced and the different components are described. Experiments and results are shown in Section 3. A conclusion and future work are presented in Section 4.

2. System Model

The processes are detailed in the following subsections. A flowchart of the system is shown in Figure (1).

2.1. Motion Detection

Presently, most segmentation methods use either temporal or spatial information in the image sequence. Three conventional approaches include optical flow [10, 11], temporal differencing [12], and background subtraction [6, 13]. Optical flow-based methods give good results even in the presence of camera motion. However they are computationally expensive. Temporal differencing is more adaptive to changes in the environment and are often used to detect global changes in the scene. However, it often produces incomplete segmentation. We introduced background subtraction in Section 1 of this paper (see Equation (1)). A key advantage of background subtraction is that it preserves the shape of the segmented objects. There are two key challenges when employing background subtraction. Training the initial model and updating the model.

2.1.1. Training

We train the initial background model using a pixel-wise temporal median filter. The background model consists of the median gray-level intensity values, for each pixel. In our system, the initial model can be learned via a training sequence if available or, it will automatically build the model using the first $\gamma$ seconds of the testing sequence.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(2)

2.1.2. Updating the background model

Most methods employ a constant learning rate for background updating [4, 6]. The problem with a constant learning rate is that objects which tend to exhibit little or no motion for an extended period of time are absorbed into the background model and the old model is forgotten. The assumption is that the most consistently observed features at each pixel originate from the background. Stauffer et al. [6] solved this problem by maintaining $K$ versions of the background, so if the once inactive object became active again, the system would revert to one of the $K$ versions that accurately modelled the current background. The problem however with this approach is that if this version of the background became the $K$th component, it would be replaced with a new component centered at the current observation with a large variance.

To highlight the magnitude of this problem, consider the scenario in an airport terminal of a person who places his/her bag on the floor somewhere inconspicuously, and walks away. A worst case scenario would be that the bag contains explosives. Although we do not want the adaptivity of the model to blend previously active objects into the background, we do want to limit their effects on tracking other objects in the same 3D space, hence avoiding occlusion.

Each pixel in the background frame is modelled as a Gaussian distribution $N(\mu, \sigma)$ with the mean $\mu$ and variance $\sigma^2$ defined in Equations (2, 3) respectively.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

(3)
foreground segmentation which we use as input into our motion
then applied as defined in Equation (5), to produce the coarse
apply Equation (1), to extract the raw motion. Binarization is
Now that we have an adaptive model of the background, we
sified as a background sample. Details on the calculation of the
distribution is updated. Hence, we employ a selective update
procedure. We add a new sample to the model only if it is clas-
s, the very recent variability that each pixel exhibits over a
period of $\varphi$ seconds. Throughout the sequence, after
every $\varphi$ seconds, the system calculates the Mahalanobis dis-
tance (see Equation (4)) between the pixel’s recent variability
distribution and the current background distribution $\omega_t(x,y)$. If
This distance is less than $2.5\sigma$ standard deviations, the current
distribution is updated. Hence, we employ a selective update
procedure. We add a new sample to the model only if it is clas-
sified as a background sample. Details on the calculation of the
dynamic Gaussian distribution will be discussed later.

$$\text{distance} = \frac{|| (I(x,y) - \mu(x,y)) ||}{\sigma} \tag{4}$$

2.2.1. Suppression Layer 1 - The Pixel Process
Now that we have an adaptive model of the background, we
apply Equation (1), to extract the raw motion. Binarization is
then applied as defined in Equation (5), to produce the coarse
foreground segmentation which we use as input into our motion
suppression component.

$$\beta_t(x,y) = \begin{cases} 
255 & \text{if } \alpha_t(x,y) > \theta \\
0 & \text{otherwise} \end{cases} \tag{5}$$

where $\beta_t(x,y)$ is the binarized image at time $t$, $\alpha_t(x,y)$ the
raw motion at time $t$ and $\theta$ the threshold. Note that $\theta$ should be
small so as to guarantee a coarse segmentation.

2.2.2. Noise Removal
The second suppression layer considers the second source of
false detections, that of random noise in the scene. Due to its
random nature, these noisy pixels are not modelled as being
part of the usual variation exhibited by each pixel. We suppress
this noise by applying a fast connected component finding al-
gorithm directly to the results of the first layer of suppression
$\text{layer1}_t(x,y)$. These connected components are represent-
ed by $C_i$, $i = 1...n$. We then threshold those components that are
of a negligible size. This process is expressed in Equation (7).

$$\text{layer2}_t(x,y) = \begin{cases} 
255 & \text{if } C_i^{Area} \geq c \text{ pixels} \\
0 & \text{otherwise} \end{cases} \tag{7}$$

where $c$ is a constant.

Note that the threshold must be effective in removing noise
without affecting the objects of interest in the scene.

2.2.3. Suppression Layer3 - Component Filtering
After the first two layers of motion suppression, we have an
image that clearly distinguishes the objects of interest as white
pixels contrasted against the black background. To avoid the
problem of a single object splitting into $n$ separate objects, a
morphological dilation operator [14] is applied to group fore-
ground pixels that are in close proximity to each other. We

$$\sigma^2 = \frac{1}{n-1} \left( \sum_{i=1}^{n} x_i - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 \right) \tag{3}$$
then apply connected component finding to merge all the foreground pixels that are connected into homogenous clusters. A final layer of component filtering is then applied to remove artifacts left-over from the previous two steps.

2.3. Features Extraction

In order to perform high-level analysis, key features need to be extracted. For each object a number of features are extracted including object centroid, area, density, velocity, radius, form factor, eccentricity as well as its contour. In addition, we create unique MBR objects (Minimum Bounding Rectangles) for each component.

2.4. Inactive Object Detection

Now that we have a clear and distinct foreground/background segmentation, we go a step further in distinguishing between active and inactive objects in the environment. As alluded to earlier, the tracking of inactive objects in security sensitive areas is of paramount importance. We are able to detect whether a previously active object has become inactive by analyzing its trajectory. In particular, we perform shift analysis of the objects centroid. The centroid was chosen as it is the most stable feature in relation to this problem. Warnings are fired by the system based on two user specified parameters namely, $\phi_{\text{min}}$, the minimum time in seconds that an object is allowed to be stationary in the scene and $\phi_{\text{max}}$, the maximum time in seconds that an object can be stationary in the scene after the object has been flagged by the system as inactive.

3. Experiments and Results

Experiments have been conducted on a variety of sequences containing both various changes in the background and complex activities of the foreground. Changes in the background include irrelevant motion arising out of scene clutter, illumination changes, shadows and random noise. Complex activities of the foreground include overstays of targets. This problem is dealt with directly by our inactive object warning module. We demonstrate further that our model is not limited to a certain position or camera geometry and in addition, can handle various environments both indoor and outdoor in an unsupervised manner.

Experiments were carried out using two publicly available data sets. The PETS$^1$ 2004 & 2006 Benchmark Data and the video footage$^2$ provided by Elgammal et al. [9].

Performance evaluation metrics are given in Table (1). Figures (2 & 3) highlight the various stages of our multi-layer extraneous motion suppression scheme. These results demonstrate the robustness of our motion detection algorithm in complex backgrounds. In Figure (4) we demonstrate the application of our inactive detection module to monitor stationary and non-stationary vehicles at traffic intersections. Figures (5 & 6) demonstrate our inactive detection module applied to detecting unattended luggage. The scenario in Figure (5) contains a single person with ski equipment who loiters before abandoning the item of luggage. Similarly, the scenario in Figure (6) contains a single person who drops her bag, circles it, and then walks away. In both cases, the abandoned luggage are detected as inactive objects.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>#Frames</th>
<th>Detection Rate</th>
<th>False Alarm Rate</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>797</td>
<td>0.94</td>
<td>0.0263</td>
<td>94%</td>
</tr>
<tr>
<td>2</td>
<td>854</td>
<td>0.7945</td>
<td>0</td>
<td>81%</td>
</tr>
<tr>
<td>3</td>
<td>1539</td>
<td>0.7725</td>
<td>0</td>
<td>77%</td>
</tr>
<tr>
<td>4</td>
<td>3400</td>
<td>0.9001</td>
<td>0.0151</td>
<td>90%</td>
</tr>
<tr>
<td>5</td>
<td>862</td>
<td>0.9164</td>
<td>0</td>
<td>92%</td>
</tr>
</tbody>
</table>

The following parameter settings were applied in our experiments. For the training of the background model, $\gamma = 10$ seconds. This parameter is applied only if there is no training sequence available. Hence, it is applied to the testing sequence and is dependent on the volume of activity in the scene during this period. The update speed of the dynamic gaussian distribution $\varphi = 10$ seconds. The raw motion binarization threshold $\theta \in [15, 50]$. In Section 2.2.1, $n = 5$, $\lambda = 10$ and $c = 10$. Finally, for our inactive object detection module, the parameters $\phi_{\text{min}} = 5$ and $\phi_{\text{max}} = 10$. The system was implemented in c++ Borland 5 and runs at 21 frames per second on a standard 2.4Ghz Pentium 4 PC with 256MB RAM for 320x240 images.

4. Conclusion and Future Works

We have designed a system that is able to reliably detect objects of interest in general indoor/outdoor scenes. In addition, the proposed system can tolerate motion clutter arising out of natural scene variations as well as gradual illumination changes. To illustrate the usefulness of this approach to real-time surveillance applications, an inactive object warning module was presented. Future extensions to the system include:

1. A module to detect a sudden global change in illumination.
2. A performance evaluation module, which will provide a wide range of evaluation metrics for quantitative assessment of the performance of visual surveillance systems. As our aim is to develop more complex action and event recognition modules within our motion detection framework, performance evaluation will prove a key component.
3. Currently, the system uses a single user defined domain dependant threshold for the final suppression layer. We are currently investigating approaches, which will eliminate the need for the user to set this threshold when setting up the system in a particular environment. This will then make our system fully automated.

In addition, we would like to investigate the development of plugins for context-controlled motion detection.

5. Acknowledgements

The authors would like to thank the National Research Foundation of South Africa and PRISM/CSIR for their financial support of this work.

6. References


---

$^1$ftp://ftp.pets.rdg.ac.uk

Figure 2: Sequence 1 - Modelling complex backgrounds (a) Original frame with bounding boxes around objects of interest (b) Raw motion (c) Layer 1 suppression and (d) Layer 2 suppression.

Figure 3: Sequence 2 - Modelling complex backgrounds (a) Original frame with bounding box around object of interest (b) Raw motion (c) Layer 1 suppression and (d) Layer 2 suppression.

Figure 4: Sequence 3 - Inactive detection module applied to road traffic analysis (a) Original frame (b) Active objects and (c) Inactive objects.

Figure 5: PETS 2006 Dataset, Sequence 4 - Inactive detection module applied to abandoned luggage detection (a)-(b) Person enters scene with luggage (c) Person abandons luggage (d) System fires warning of abandoned luggage - Green bounding box (e) System fires second warning indicating overstay of target - Red Bounding Box (f) Inactive object still detected as foreground.
Figure 6: PETS 2004 Dataset, Sequence 5 - Inactive detection module applied to abandoned luggage detection (a)-(b) Person enters scene with luggage (c) System fires warning of abandoned luggage - Green bounding box (d) System fires second warning indicating overstay of target - Red bounding box


