# Comparison of Background Models for Video Surveillance

Matthew Fettke
*Flinders University*
*School of Informatics and Engineering*
*Matthew.Fettke@flinders.edu.au*

Matthew Naylor
*Vision Systems Limited*
*Matthew.Naylor@vsl.com.au*

Karl Sammut
*Flinders University*
*School of Informatics and Engineering*
*Karl.Sammut@flinders.edu.au*

Fangpo He
*Flinders University*
*School of Informatics and Engineering*
*Fangpo.He@flinders.edu.au*

## Abstract

*Background modelling is a common form of motion detection employed by many autonomous video surveillance systems. Accurately modelling the background is a challenging task, particularly for outdoor scenes where factors such as background motion and camera shake can cause the mistaken detection of foreground objects. Recent research has developed background models that are capable of detecting foreground motion in real-time while ignoring most of the background motion, but it is not clear how well these models would perform on outdoor scenes that exhibit typical video surveillance problems. The aim of this paper is to assess the performance of leading background models (namely $W^4$, the Hybrid Detection Algorithm, and Three-frame Temporal Difference), using video sequences that contain problems which trouble existing video surveillance systems. The strengths and weaknesses of these background models are reported and analysed, with the aim of identifying suitable directions for the development of robust background models for motion detection in outdoor video surveillance systems.*

## 1. Introduction

When monitoring a scene for surveillance purposes, it is often desirable to distinguish between objects that move normally, such as trees blowing in the wind, and objects that move in a way that is unusual, such as a person walking in an area that is forbidden. The normally moving objects can be referred to as background objects, while the objects with unusual movement can be considered as foreground objects. Accurately classifying these two types of objects for a given video sequence is a challenging task for an automatic video motion detection system, but is crucial for automating video surveillance. One approach is to model the background [1].

In background modelling, the video sequence is analysed and a model of the scene is constructed over time to represent the usual scene content. New frames of the video sequence can then be compared to this model, with regions that differ being classified as foreground motion. However, if this technique is based on pixel intensities, the differences may not be solely due to foreground motion. Lighting changes, noise, and camera movement all cause intensities to change, and can cause false classification. Two common sources of false detection are background movement and camera shake [1, 2, 3, 4]. Both of these are are frequently encountered in outdoor video surveillance systems.

Modern surveillance systems expect the motion detection system to accurately classify groups of detections as entire objects. It must also be able to operate in real-time on minimal hardware using at least CIF resolution images. A motion detection system based on background modelling is able to meet these criteria, and there are many examples of such systems.

The temporal difference technique [1] uses the previous frame as the background model. Motion is detected using the differences between pixel intensities - if the magnitudes of the differences are greater than a threshold, the pixel is classified as foreground. The main problem with this method is that it detects all the motion within the scene, leading to the development of more advanced models of the background. However, many are based on assumptions that make them fundamentally unsuitable for use in an outdoor surveillance system, with two examples being Pfinder [5] and Wallflower [3]. Pfinder is unable to adequately model typical outdoor background movement having been designed with a different purpose in mind, and Wallflower is not able to properly model tree movement - see [6] for a more complete description of their respective shortcomings.

This paper aims to compare the background models known as $W^4$ [7], the Hybrid Detection Algorithm (HDA) [8], and the Three-frame Temporal Difference (TTD) to identify the advantages of developing an advanced model of the background. Video sequences representative of typical outdoor surveillance scenes are used to evaluate the background models, trying to identify the model most appropriate for outdoor video surveillance. Although [3] includes a review of some recently developed background models, it focusses more on problems encountered in indoor surveillance. The contribution of this work is to concentrate on some typical outdoor surveillance problems - background movement and camera shake.

The three background models to be implemented are described in Section 2. The experiments performed in this paper are outlined in Section 3, and the results of these experiments are analysed in Section 4.

## 2. Background models

Although descriptions of the three background models to be compared in this paper have been previously provided [8, 7, 6], they are given here in order to allow true replication of the experiments. All three background models have parameters that are crucial to their performance, so they need to be stated clearly.

### 2.1. Three-frame Temporal Difference

The temporal difference technique is based on the assumption that any interframe change of pixel intensities is the direct result of motion within the scene. The Three-frame Temporal Difference is a variation designed to reduce the effects of noise, and uses the previous two frames as the background model. Every pixel $x$ in frame $I_t$ is compared to that in frames $I_{t-1}$ and $I_{t-2}$. If the magnitude of either difference is less than a preset threshold, $Th$, then that pixel is labelled as background:

$$\begin{gathered}(|I_{t-1}(x) - I_t(x)| > Th) \\ \text{AND} \\ (|I_{t-2}(x) - I_t(x)| > Th).\end{gathered} \tag{1}$$

To eliminate noise further, objects made up of three pixels or less are removed from the final foreground motion mask. The outcomes of a preliminary investigation indicate that a value of 20 for $Th$ gives the best results.

There are two main problems [1] with the temporal differencing technique. If a homogeneous object undergoes motion, the temporal difference fails to detect that interior pixels have moved. The only pixels detected are those at the "wavefront" of the motion. Thus, further processing must be performed to recover the entire moving object. The second problem is that once an object stops moving, it is no longer detected. This is obviously unsatisfactory because it is still important to detect foreground objects when stationary. In reality, motion detection is rarely used by itself for foreground object detection; however, an ability by the motion detector to detect objects that are stationary for small periods of time is still extremely beneficial.

### 2.2. Hybrid Detection Algorithm

The HDA is described in [8] and incorporates the three-frame temporal difference technique to detect all pixels that *may* contain motion. The background model consists of two adaptive parameters for every pixel $x$ in image $I_t$ - the current background intensity $B_t(x)$ and the threshold $T(x)$. Both are updated depending on whether the pixel $x$ is determined to be static or moving:

$$B_{t+1}(x) = \begin{cases} \gamma B_t(x) + (1-\gamma)I_t(x) & x \text{ static;} \\ B_t(x) & x \text{ moving.} \end{cases} \tag{2}$$

$$T_{t+1}(x) = \begin{cases} \gamma T_t(x) + \\ 5(1-\gamma)(|I_t(x) - B_t(x)|) & x \text{ static;} \\ T_t(x) & x \text{ moving.} \end{cases} \tag{3}$$

A three-frame difference algorithm (1) is used to determine the subset of pixels in the image $I_t$ that might be moving, that is, if the differences are greater than the threshold then pixel $I_t(x)$ is moving. The resulting regions of pixels then undergo two morphological dilations followed by one erosion. Although this technique is not described in [8] it ensures that pixels are added to larger objects if they are relatively close, thus reducing the number of detected objects and improving the quality of those objects detected. Connected component clustering is performed, providing each object with a unique label.

Background subtraction is then performed on the bounding box of each region $R_n$. Each foreground object $b_n$ is determined by comparing the pixel intensities within $R_n$ with the background intensity. That is,

$$b_n(x) = (x : |I_n(x) - B_n(x)| > T_n(x), x \in R_n(x)). \tag{4}$$

The initial background intensity, $B_0(x)$ is the same as the pixel values in the second image, $I_1(x)$. The initial thresholds, $T_0(x)$, are set to 20 based on the outcomes of an initial investigation. The value of $\gamma$ is not provided in [8], so is based on experimentation. A value of $\gamma = 0.995$ is used in this paper since this is the smallest value that prevents slowly moving foreground objects, such as a person walking, from being incorporated into the background model.

Collins *et al.* [8] report that this technique is very fast and is used as their primary motion detection method. However there are some basic problems with the technique as described. As it uses temporal differencing to determine the regions on which to perform background subtraction, moving objects that become stationary will not be detected. Moving objects that are detected are not guaranteed to be detected as one whole object. That is, as the pixels detected by the temporal difference step may only be a subset of the whole object, they may be spatially separated such that the dilation process does not bring them together. The connected component algorithm will then assign the same object multiple labels, which may lead to confusion in a higher level process.

### 2.3. $W^4$

$W^4$ [7] is a real-time system designed to detect multiple people in an outdoor environment. It is able to operate at 25 frames per second using $320 \times 240$ sized greyscale images on a dual 300 MHz Pentium II PC. The background model used represents each pixel $x$ with three values - the minimum pixel value $m(x)$, the maximum pixel value $n(x)$, and the maximum difference of pixel intensity between consecutive frames $d(x)$ observed during a learning phase.

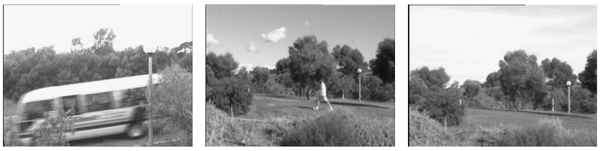
Foreground objects in the scene are detected by comparing the current frame to the background model. If

$$(|I_t(x) - m(x)| > d(x)) \quad \text{AND} \quad (|I_t(x) - n(x)| > d(x)) \tag{5}$$

then pixel $I_t(x)$ is classified as foreground; otherwise, $I_t(x)$ is classified as background. A connected components algorithm is then applied to the resulting motion mask, with all objects smaller than 3 pixels assumed to be noise.

## 3. Experiments

Each background modelling technique will be evaluated using seven video sequences recorded in our laboratory. While these sequences are not standard test data, they are typical of outdoor scenes and exhibit the problems of background motion and camera shake. Each sequence was recorded at 25 frames per second and is made up of $384 \times 288$ sized frames. All frames are comprised of 8-bit greyscale pixels, typical of the frames used in surveillance applications. A brief description of the sequences follows, and example frames from the sequence set are shown in Figure 1.



(a) bus (b) walk (c) shake

**Figure 1. Examples frames from the sequences used during the experiments described in this paper.**

**bus:** This sequence is comprised of 740 frames and contains quite substantial tree movement in the foreground of the scene. A bus enters the field of view at frame 640 and takes 50 frames to traverse the scene. The bus is almost $\frac{1}{3}$ of the image size.

**walk:** This sequence contains 988 frames and contains much less background movement, although a light breeze causes the trees to move slightly. A person walks across the scene from frame 700 until frame 935.

**shake:** This sequence is made up of 990 frames and does not involve any foreground objects. The camera undergoes quite substantial shake similar to that caused by gusts of wind.

**collide:** This sequence contains 990 frames and exhibits small amounts of background movement caused by a slight breeze. Two people enter the scene by frame 700 and leave by frame 880.

**overtake:** Two people take 200 frames to move through the scene that contains background movement caused by a light wind.

**disperse:** A group of four people are stationary for 640 frames before dispersing from the scene by frame 840.

**approach:** Two people are stationary for the entire sequence, while a third person enters the scene after frame 700. A strong breeze causes quite substantial tree movement in the background of the scene.

These sequences are used as the basis for three experiments to evaluate the accuracy of the background modelling techniques. Each experiment involves obtaining a motion mask at various stages of the sequences, which is then compared to the hand-generated ideal motion mask using the bit-wise exclusive-OR operator. The differences are described in terms of false positives and false negatives. This comparison is rather basic in nature, and provides results that would be misleading if interpreted in isolation. However, when used in conjunction with the qualitative approach of comparing the resultant images, it proves to be quite adequate for the purposes of this evaluation.

### 3.1. Experiment 1 - Background motion removal

The ability of each technique to represent the background motion in scenes from a stationary camera under constant illumination is examined. The first 625 frames of the bus sequence and the first 640 frames of the approach sequence contain large amounts of background motion, while the first 675 frames of the walk sequence, 650 frames of the collide sequence, 660 frames of the overtake sequence, and 610 frames of the disperse sequence contain minimal background motion. Each technique is used to produce a motion mask at the end of these sections of the sequences.

### 3.2. Experiment 2 - Foreground motion detection

The ability of each technique to allow reliable detection of foreground objects in the presence of background motion is examined. The bus sequence contains a large foreground object, while the walk, collide, overtake, disperse, and approach sequences contain small foreground objects. Each technique is used to produce a motion mask at frame 665 of the bus sequence, frame 800 of the walk sequence, frame 770 of the collide sequence, frame 830 of the overtake sequence, frame 700 of the disperse sequence, and frame 730 of the approach sequence.

### 3.3. Experiment 3 - Camera shake

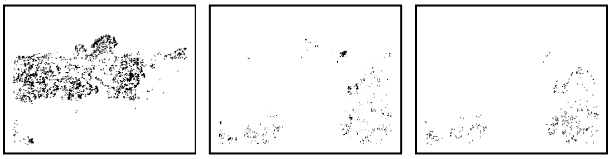
The ability of each technique to remove the effects of camera shake under constant illumination is examined. Each technique is used to produce a motion mask at frames 875 and 900 of the shake sequence.

## 4. Analysis

The analysis of the results involves obtaining the total number of erroneously classified pixels and comparing the motion masks produced by each technique. Any deficiencies are then explained, and the best technique for each scenario identified. The results of the experiments are summarised in Table 1.

### 4.1. Experiment 1 - Background motion removal

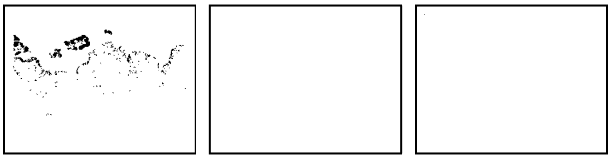
Experiment 1 shows that the TTD is the technique best able to remove the effects of wind in a scene. It can be seen in Figure 2 that the $W^4$ method detects much more of the background motion than the other two techniques, producing more than 10 times the number of incorrectly classified pixels. Further analysis shows that the learning phase of the algorithm is the cause of these mis-classifications. If the intensity of pixel $I_t(x)$ has a large range of values but only changes slowly, there will be a large difference between $m(x)$ and $n(x)$, but $d(x)$ will be small. It is therefore likely that $I_t(x)$ differs from $m(x)$ and/or $n(x)$ by an amount much larger than $d(x)$, resulting in the possibility that $I_t(x)$ is wrongly classified as foreground (see (5)). Examples of this can be observed in Figures 2 and 3 where clouds and trees are detected as foreground objects by the $W^4$ technique while both the HDA and TTD have correctly classified them as background.



(a) $W^4$ (b) HDA (c) TTD

**Figure 2. The motion masks produced after the first 625 frames of the bus sequence show that the TTD is able to model more of the tree movement than $W^4$ and the HDA.**

Figure 3 highlights the advantage of using the previous two frames as the background model. The walk sequence contains only a small amount of background motion, so the interframe pixel intensity differences are small. The TTD does not detect much, if any, of the background motion as a result. As the HDA algorithm uses the TTD as the initial phase, it also detects very little of the background motion.

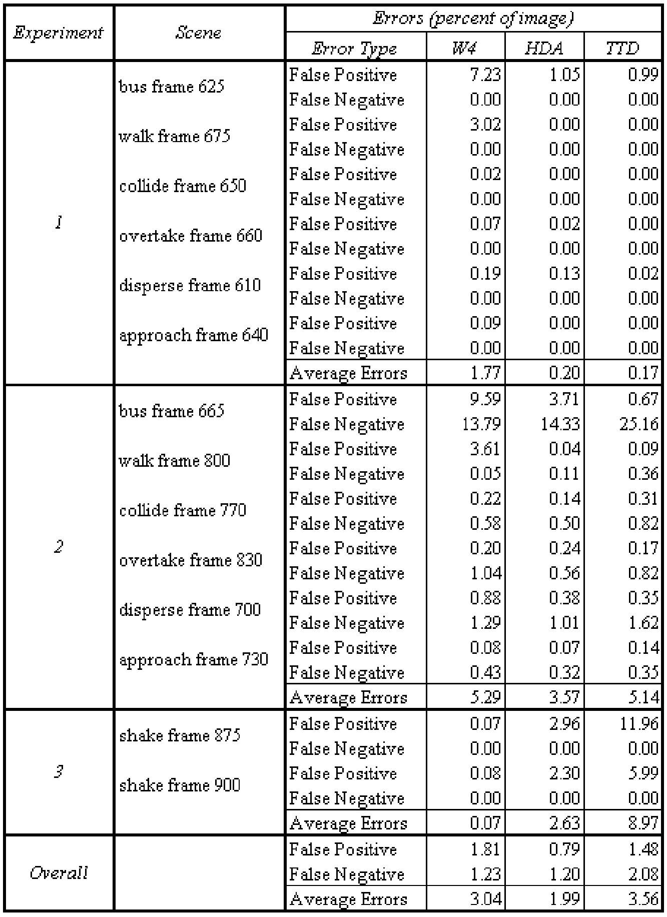


(a) $W^4$ (b) HDA (c) TTD

**Figure 3. The motion masks produced after the first 675 frames of the walk sequence show that the $W^4$ model does not represent the background movement well, while both the TTD and HDA techniques do not detect any background movement.**

The results from Experiment 1 show that the TTD is slightly more robust to the background motion contained in the sequences than the HDA, while $W^4$ is prone to mis-

**Table 1. Experimental results.**



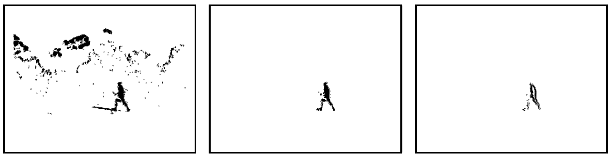
| *Experiment* | *Scene* | *Errors (percent of image)* | | | |
|---|---|---|---|---|---|
| | | *Error Type* | *W4* | *HDA* | *TTD* |
| *1* | bus frame 625 | False Positive | 7.23 | 1.05 | 0.99 |
| | | False Negative | 0.00 | 0.00 | 0.00 |
| | walk frame 675 | False Positive | 3.02 | 0.00 | 0.00 |
| | | False Negative | 0.00 | 0.00 | 0.00 |
| | collide frame 650 | False Positive | 0.02 | 0.00 | 0.00 |
| | | False Negative | 0.00 | 0.00 | 0.00 |
| | overtake frame 660 | False Positive | 0.07 | 0.02 | 0.00 |
| | | False Negative | 0.00 | 0.00 | 0.00 |
| | disperse frame 610 | False Positive | 0.19 | 0.13 | 0.02 |
| | | False Negative | 0.00 | 0.00 | 0.00 |
| | approach frame 640 | False Positive | 0.09 | 0.00 | 0.00 |
| | | False Negative | 0.00 | 0.00 | 0.00 |
| | | Average Errors | 1.77 | 0.20 | 0.17 |
| *2* | bus frame 665 | False Positive | 9.59 | 3.71 | 0.67 |
| | | False Negative | 13.79 | 14.33 | 25.16 |
| | walk frame 800 | False Positive | 3.61 | 0.04 | 0.09 |
| | | False Negative | 0.05 | 0.11 | 0.36 |
| | collide frame 770 | False Positive | 0.22 | 0.14 | 0.31 |
| | | False Negative | 0.58 | 0.50 | 0.82 |
| | overtake frame 830 | False Positive | 0.20 | 0.24 | 0.17 |
| | | False Negative | 1.04 | 0.56 | 0.82 |
| | disperse frame 700 | False Positive | 0.88 | 0.38 | 0.35 |
| | | False Negative | 1.29 | 1.01 | 1.62 |
| | approach frame 730 | False Positive | 0.08 | 0.07 | 0.14 |
| | | False Negative | 0.43 | 0.32 | 0.35 |
| | | Average Errors | 5.29 | 3.57 | 5.14 |
| *3* | shake frame 875 | False Positive | 0.07 | 2.96 | 11.96 |
| | | False Negative | 0.00 | 0.00 | 0.00 |
| | shake frame 900 | False Positive | 0.08 | 2.30 | 5.99 |
| | | False Negative | 0.00 | 0.00 | 0.00 |
| | | Average Errors | 0.07 | 2.63 | 8.97 |
| *Overall* | | False Positive | 1.81 | 0.79 | 1.48 |
| | | False Negative | 1.23 | 1.20 | 2.08 |
| | | Average Errors | 3.04 | 1.99 | 3.56 |

classifying many of the pixels containing background motion.

### 4.2. Experiment 2 - Foreground motion detection

The results from Experiment 2 show that the HDA is better able to properly detect the foreground objects in the video sequences. The TTD removes more of the background motion in the bus sequence than the other two techniques, but fails to detect most of the interior pixels of the bus object. This can also be seen in Figure 4 where the interior of the person is not detected by the TTD as undergoing motion. Although the $W^4$ technique has the lowest number of false negatives, it detects an extremely high number of false positives as discussed in the previous section.

The large number of false negatives produced by the three techniques for the bus sequence is due to the fact that the ideal motion mask quite correctly does not include the occluding leaves. The three techniques have shown that they are able to ignore much of the motion associated with tree movement, and so have not detected many of the tree regions as foreground movement. The areas where the ignored motion corresponds to the bus object have contributed to the very large false negative value.

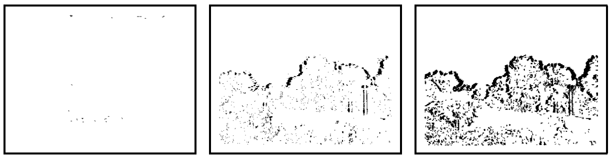


(a) $W^4$ (b) HDA (c) TTD

**Figure 4. The HDA is the most successful technique at extracting the entire person as foreground movement while not detecting any of the tree motion.**

### 4.3. Experiment 3 - Camera shake

Experiment 3 shows that the $W^4$ technique is the best at ignoring the motion caused by camera shake, highlighted by the motion masks shown in Figure 5. Both the TTD and HDA techniques detect much more motion because the interframe camera movement is quite large, meaning that the previous two frames are often very different to the current frame.



(a) $W^4$ (b) HDA (c) TTD

**Figure 5. Of the three techniques, only $W^4$ is able to model the shaking of the camera, while both the TTD and HDA produces large numbers of false positives.**

Most video surveillance systems would use an image stabilisation routine to remove the effects of camera shake. However, this solution is not perfect, as it is sometimes impossible to perform perfect stabilisation [9]. Therefore, it is an advantage if the background model is able to represent the camera motion as it improves the accuracy of the motion detection algorithm.

## 5. Conclusion

Background modelling is a motion detection technique applicable to video surveillance systems as it can be implemented in real-time, is accurate, and can operate on the types of images typically used in the application. It is important for the background modelling technique employed to be able to ignore the background motion typically encountered in outdoor scenes while still being able to detect foreground objects.

This paper has evaluated two recently developed background models - $W^4$ and the Hybrid Detection Algorithm - and the more established Three-frame Temporal Difference. The two newer models greatly reduce the amount of background motion detected in an outdoor scene containing confusing phenomena compared to the simpler TTD. It was shown that the HDA is better able to model common background movement in outdoor scenes while the $W^4$ method can remove the effects of significant camera shake. Both the HDA and $W^4$ are able to detect entire foreground objects whereas the TTD failed to detect interior pixels as moving. This paper has demonstrated that the use of advanced background models allows for a reduction in the number of erroneously classified pixels.

## 6. Acknowledgment

The authors would like to acknowledge the significant contribution of Vision Systems Limited to the research presented in this paper.

## References

[1] P.L. Rosin and T. Ellis, "Image difference threshold strategies and shadow detection," in *Proceedings of the 6th British Machine Vision Conference*, 1995.

[2] Chris Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, 1999.

[3] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and Practice of Background Maintenance," in *International Conference on Computer Vision*, September 1999.

[4] J-M. Letang, P. Bouthemy, and V. Rebuffel, "Robust motion detection with temporal decomposition and statistical regularization," Tech. Rep. 2717, INRIA, 1995.

[5] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, July 1997.

[6] M. Fettke, K. Sammut, M. Naylor, and F. He, "Evaluation of Motion Detection Techniques for Video Surveillance," in *Information, Decision and Control*, 2002.

[7] I. Haritaoglu, D. Harwood, and L.S. Davis, "$W^4$: Who? When? Where? What? A Real Time System for Detecting and Tracking People," in *International Conference on Face and Gesture Recognition*, April 14-16 1998.

[8] R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson, "A System for Video Surveillance and Monitoring," Tech. Rep., The Robotics Institute, Carnegie Mellon University, 2000.

[9] I. Cohen and G. Medioni, "Detecting and Tracking Moving Objects for Video Surveillance," in *IEEE Proceedings on Computer Vision and Pattern Recognition*, June 23-25 1999.