Video Motion Detection Beyond Reasonable Doubt

Zhuo Xiao Amirsaman Poursoltanmohammadi Matthew Sorell
School of Electrical and Electronic Engineering
University of Adelaide SA 5005
AUSTRALIA
+618 8303 3226
matthew.sorell@adelaide.edu.au

ABSTRACT
We consider the analysis of surveillance video footage containing occasional activities of potential interest interspersed with long periods of no motion. Such evidence is problematic for three reasons: firstly, it takes up a great deal of storage capacity with little evidential value; secondly, human review of such surveillance is extremely time-consuming and subject to errors due to fatigue; and thirdly, there is often a need to prove to the satisfaction of the Court that excised footage contains no images of evidential value. We are therefore concerned with objective, reliable detection of video motion to automate the extraction of activities of interest and to provide simple but reliable measurements to the court to prove that this is a complete record of all activities in the footage. Early results indicate that average luminance-based detection is particularly reliable, and we provide a comparison with other frame-difference techniques.

Categories and Subject Descriptors
I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – *Motion*; I.4.9 Applications; I.5.1 [**Pattern Recognition**]: Models – *Statistical*

General Terms
Algorithms, Measurement

Keywords
video motion detection, forensics, surveillance, law enforcement, evidence

1. INTRODUCTION
Consider a common scenario in which a surveillance camera is used to record days or weeks of video footage of, for example, the entrance to a building, for the purpose of gathering evidence in the context of a criminal investigation. The camera is stationary, and most of the time the scene is deserted, with occasional activity lasting seconds or minutes.

e-Forensics 2008, January 21-23, 2008, Adelaide, Australia.

The usual practice is for such footage to be reviewed by an operator who extracts events of interest manually. It is common, then, for the original footage to be discarded so that video tapes or hard-drives can be re-used. This approach may lead to difficulties in having the evidence accepted in Court at a later date. In particular, if only selected footage is shown to the Court, and the original, full, footage is not available, it would be possible for the defence to have that evidence rejected on the basis that it is incomplete and does not show the relevant hypothetical evidence of the defendent's innocence.

The purpose of our research is to provide a technical solution to the challenges raised in this scenario. We propose to process the footage by one or more motion detection algorithms with well-defined statistical properties, and hence

1. Automate the first-stage process of motion footage extraction and
2. Provide a graphical record of motion measurements which can demonstrate that all motion events have been preserved as evidence, with a quantifiable probability of error.

To date, almost all research in the general area of motion detection has assumed that the video frames contain some form of movement, and the focus has been the classification of that movement into matters of interest, ranging from the identification of moving objects [7], to discriminating between background objects waving versus people walking into a scene [8], to identifying suspicious behavior by individuals within a scene [4, 5, 11]. Such research fits within the realm of artificial intelligence. Automated motion detection within an area of interest is commonly implemented using a passive infrared sensor (PIR), but this approach offers no guarantee that all motion has been captured, and indeed there are well-established techniques for thwarting such a sensor, for example by moving slowly behind a thermal screen such as a blanket.

2. PROPOSED APPROACH
We envisage a system consisting of a means of digital video capture, real-time analysis and storage of results. The video capture might consist of a camera connected directly to a processor on location, a capture card for transfer from, for example, video tape, or a digital video file. There would be three significant output streams from the system:

1. A regular video frame at a very slow rate (for example one frame every ten seconds) to provide evidence that the camera is working properly. This would also allow fast, start-to-

finish review of the complete content of the footage, albeit without temporal detail.

2. Whenever motion is detected, that footage, with an appropriate lead in and lead out to provide context, at the full frame rate.

3. A timeline containing one or more motion detection metrics, so that it is possible to review and demonstrate to a Court that footage which has not been captured does not contain motion which might be of interest.

The physical architecture, and the first two output streams, are straightforward to implement and are not of significant research interest. However the third requirement has, surprisingly, not been adequately addressed in the literature.

3. PROBLEM FORMULATION

We consider the situation in which we have a sequence of video of length N frames F at some arbitrary resolution $x_{\max}$ by $y_{\max}$. Within each frame we define the Region of Interest as an arbitrary window W, which could include the entire frame, defined by the indicator function $I_{ROI}(i,j)$. We are concerned with identifying changes from one window to the next which indicate motion.

$$W(k,i,j) = F(k,i,j)I_{ROI}(i,j)$$

where

$k = 1,..,N$ is the frame number

$i = 1,..,x_{\max}$ and $j = 1,..,y_{\max}$ is the pixel location

$$I_{ROI}(i,j) = \begin{cases} 0 & \text{if } (i,j) \notin W \\ 1 & \text{if } (i,j) \in W \end{cases}$$

A motion detection test T is defined as an arbitrary motion metric between the window at time k and the windows leading up to time k. This test is compared with some threshold τ. This generic formulation allows for comparison with, for example, the immediate previous window, or a fixed or varying estimate of the background. If the test T exceeds the threshold τ, we say that there is evidence to support the hypothesis that a motion event has been detected.

$$T(W(k) \mid W(1),..,W(k-1)) \underset{H_0}{\overset{H_1}{\gtrless}} \tau$$

where

H_0 is the hypothesis that there is no motion event

H_1 is the hypothesis that there is a motion event

As it is the intention of the algorithm to capture video sequences which include motion events, it is assumed that such capture would include a lead-in and lead-out sequence of M frames to provide context. A True Detection event, generally defined as the case where a motion event is detected given that a motion event has actually occurred, is specifically defined as the occurrence of a true motion event within M frames of motion being detected in frame k. This definition allows for multiple detections within a short space of time to generate a valid continuous sequence which includes a true motion event (which would usually take place over multiple frames). A Miss event is similarly defined as the situation in which a true motion event occurs at frame k but there is no motion detected by the detection test within the timeframe window of M frames before and after the frame k.

Conversely a False Alarm is specifically defined as a motion detection at frame k for the situation in which no true motion event is identified within the timeframe window of M frames before and after the frame k.

We formulate our detection requirements according to the Neyman-Pearson criterion [9] in a slightly non-conventional way. We constrain the Probability of Miss to be less than some nominal probability β, and then seek to minimise the Probability of False Alarm. In practice, such optimisation can only be performed empirically across a range of sample video footage, and by then comparing performance of a range of motion detection metrics. It should also be noted that $T(k)$ is logged to demonstrate when motion was, and was not, detected, for possible presentation as evidence.

4. CHANGES IN VIDEO SEQUENCES

Restricting the attention of this research to the scenario in which the camera is fixed and the scene is largely static significantly simplifies analysis. However, the situation is complicated by the range of changes which can occur within a video sequence. Although we are not concerned with the automatic classification of a motion event, it is nevertheless necessary to consider changes which are clearly not motion related versus those which might indicate motion event of interest.

4.1 Changes not indicative of motion

There are three types of changes of interest in the first category:

1. Imaging noise
2. Slow changes in lighting dependent on the time of day
3. Background random or periodic motion such as trees waving in the wind

4.1.1 Imaging Noise

The phenomenon of so-called *salt and pepper* noise in video images is well understood. Its occurrence is largely due to silicon imaging sensors being sensitive to temperature-dependent fluctuations in electron motion, and is particularly noticeable in low-light situations.

There are other sources of similar random or pseudo-random noise, including noise introduced by storage and retrieval from video tape, noise introduced by analog transmission of a video signal, compression artifacts in a digitally compressed video sequence, and digital noise introduced by errors in the transmission of a digital video signal.

Another source of noise, or more accurately distortion, may occur when a video signal is transcoded, especially in the analog to digital conversion process. When the frame rate of the incoming signal does not match the coding frame rate, it is common for the capture algorithm to skip or repeat frames as needed to match the frame rates. Although this is easily identified through the detection of periodic change effects, it is an unwelcome source of artefacts and it is advisable for the capture process to ensure that each frame is captured only once.

For the purpose of these results, we consider that sensor noise is the dominant form of imaging noise, which is well modelled as an additive Gaussian signal in the distribution of luminance with

some estimatible variance. This effect therefore dominates in setting the threshold for detection, as too small a threshold will result in false alarms due to random sensor fluctuations.

4.1.2 Lighting Changes

Sunlight plays a significant part in any surveillance footage, even when artificial lighting is used indoors. In addition to the variation throughout the day as the sun moves overhead, it should also be noted that clouds can cause shadows which might be interpreted as a moving object. In the case of our analysis, we assume that natural lighting accounts for a slow variation in shadowing and overall luminance throughout the day. We allow for such changes by tracking and re-centering the mean of our motion metric over time, where appropriate.

However, there are some lighting changes which are properly considered to be motion-like events, as discussed in Subsection 4.2.2.

4.1.3 Background Random Motion

It is also desirable to ignore background random or periodic motion. The most obvious example of such motion is trees waving in the background, but any motion at a sufficient distance to be out of scope of interest may be considered in this way.

There are three common ways of dealing with this type of motion:

1. Excise the region in which such motion occurs, which we have allowed for by specifying a region of interest,
2. Where appropriate, treat such motion as random noise and model the motion as part of the imaging noise, which we can achieve by changing our detection threshold, or
3. Classify specific moving objects and isolate them from consideration. We do not consider this approach, but note that it is the subject of a large body of research, such as [8].

4.2 Motion events

Events which are likely to be of interest fall into two general categories: objects which are moving within the video sequence, and lighting changes of interest.

4.2.1 Moving Objects

Moving objects include not only people walking into a Region of Interest, but might also include small animals and other objects. For the purpose of establishing whether an object has moved within the scene, it is important that we be able to detect such movement. The classification of whether that movement is of evidential interest, or clearly of no interest, is a matter which can be decided relatively efficiently by a human operator. However, it should be noted that even if a cat walks through the scene, it is important to retain this footage should the motion event be questioned in Court.

4.2.2 Fast Lighting Changes

Fast lighting changes do not necessarily indicate motion, but they will inevitably be detected by a motion detection algorithm and will often be associated with motion of interest. Examples include artificial lights being switched on or off, flashes due to reflections from cars driving past, and conceivably gunshot or camera flash events. It should also be noted that cameras with automatic brightness/contrast settings may trigger a sensor if there is a large step adjustment in the camera settings, although this will usually occur in response to a change in conditions due to a legitimate motion event.

5. MOTION DETECTION ALGORITHMS

In our experimental approach we consider four early candidates for motion detection:

1. Sum of Absolute Differences
2. Maximum Macroblock Sum of Absolute Differences
3. Average Luminance
4. Independent Luminance Entropy

5.1 Sum of Absolute Differences

The Sum of Absolute Differences (SAD) metric is well established as a technique for motion estimation in video compression [1]. Its performance is inferior, but close to, the use of the mean-squared error, except that the number of processor instructions is substantially less. The metric compares the current frame with the previous frame as a reference, but can be modified to consider the current frame against an estimate of the background.

$$T_{SAD}(k) = \sum_{j=1}^{y_{\max}} \sum_{i=1}^{x_{\max}} \left| W(k,i,j) - W(k-1,i,j) \right|$$

SAD has the advantages that it is computationally very simple and established in motion estimation. However, motion estimation is not the same as motion detection. In video compression, the aim of motion estimation is to identify a block of pixels from a previous reference frame which is sufficiently close to the block of interest that it can be used as an approximation for that block. There is no requirement for that match to actually refer to the same object within the scene, which means that motion vectors created using this approach do not necessarily imply movement. For this reason, even though absolute difference has been commonly proposed as a motion detection test (see for example [7]), the efficacy of this approach is questionable.

5.2 Maximum Macroblock Sum of Absolute Differences

A refinement of the Sum of Absolute Differences is to divide each window into macroblocks, as per conventional motion estimation for video compression. The window is tesselated into so-called macroblocks of nominal size (normally a square of sixteen by sixteen pixels) and the Sum of Absolute Differences is calculated for each macroblock. In conventional compression by motion estimation, each macroblock is then matched through a search algorithm with a good match from the previous frame. In our case, we compare SAD values with the macroblock in the same location in the previous frame, and use the macroblock with the largest difference as our test statistic.

$$T_{MM}(k) = \max_{u,v} \sum_{j=1}^{16} \sum_{i=1}^{16} \left| W(k, i+16u, j+16v) - W(k-1, i+16u, j+16v) \right|$$

where

$$u \in \{1,..,\lceil x_{\max}/16 \rceil\}$$
$$v = \{1,..,\lceil y_{\max}/16 \rceil\}$$

and special conditions must be considered when the window dimensions are not a rectangular multiple of 16.

It might be considered that the Maximum Macroblock SAD would perform better, as it draws attention to the region within the window with the largest change. However, while this approach would appear to offer an improvement over the full-window SAD test, it is still subject to the same impact of sensor noise.

5.3 Average Luminance

A very simple approach is to simply calculate the average luminance of the window and compare this over time. The difficulty with this approach is that simply calculating the difference in luminance from one window to the next introduces susceptibility to masking by moving very slowly. It is therefore necessary to keep track of the average luminance by applying a tracking filter of luminance on the window history, and comparing the current luminance with that expected by the tracking estimate. In this way, both step changes and slow (but not very slow) changes can be detected.

$$T_{AL}(k) = \left| \hat{Y}(k) - \sum_{j=1}^{y_{\max}} \sum_{i=1}^{x_{\max}} W(k,i,j) \right|$$

where $\hat{Y}(k)$ is the tracked estimate of the average luminance.

5.4 Independent Luminance Entropy

Entropy, in simple terms, is a measure of the order, or disorder, of the distribution of a random variable. If the luminance of each pixel is considered to be a random number drawn independently from some distribution, then the minimum entropy will be achieved if each pixel has precisely the same value. The maximum entropy will be achieved if the pixel luminance is distributed uniformly across the full range from full black (luminance = 0) to full white (luminance = 255 on an 8-bit scale). This is a crude model, because it assumes that the luminance of a given pixel has no correlation with its neighbours, but the same argument applies to the related use of entropy as the basis for information-theoretic data compression, where it is used successfully. Hence, by tracking entropy in the same way as that proposed for average luminance, changes indicate a variation in the statistical distribution of the image, and thus provide a test for motion

$$T_H(k) = \left| \hat{H}(k) - \sum_{i=1}^{255} \tilde{p}(k,i) \log\left(\frac{1}{\tilde{p}(k,i)} \right) \right|$$

where $\tilde{p}(k,i)$ is the empirical probability of the luminance taking value i in the range of 0 to 255, at time k, as determined by a normalised histogram of the luminance values. Entropy has been proposed as a metric for image segmentation in [6, 8, 10], and for the detection of scene boundaries in [2]. although that work does not consider the burden of proof required for our application.

6. EXPERIMENTAL RESULTS

A simple experiment was conducted to test the efficacy of the four proposed approaches under realistic indoor surveillance conditions.

A thirty-minute sequence of video frames was captured using a USB-connected colour camera with a resolution of 352 by 288 pixels, at a rate of 5 frames per second. Such inexpensive cameras are commonly used on this university campus for security applications. Each frame was recorded as a lossless TIF file with 8-bit resolution for Red, Green and Blue. Analysis was performed by converting each frame image to Luminance only – Chrominance information was discarded.

The scene used a combination of fluorescent lighting and shaded natural sunlight in an empty office. Movement consisted of at least one person opening the door and entering the office on two occasions, as well as shadows of people passing the office appearing on the frosted glass which was in view. A total of nine distinct real motion events were recorded, occupying a total of three minutes (180 seconds), including ten frames (two seconds) each of lead-in and lead-out.

The results of the four motion metrics are given in Figures 1 to 4. A casual glance at these figures shows that Sum of Absolute Difference-based detection techniques are much more sensitive to noise than the Average Luminance and Luminance Entropy techniques. The most likely explanation for this difference is that SAD tends to magnify noise differences, whereas the other techniques average out the noise.

For each metric, a detection threshold was set to a level which would just detect the nine true motion events. For the detection to be valid, there had to be at least one detection within the true event period, including two seconds of lead in and out. In fact, it was not possible to detect one event using SAD without setting the threshold so low as to detect motion continuously.

The detection events are also summarised in Figures 1 to 4 and are shown in comparison to the nine true motion events. It can be seen that Entropy and Luminance have a low false alarm rate, whereas false alarms are pervasive in the SAD-based detectors.

The performance of the detectors is summarised in Table 1. It can be seen that in reviewing thirty minutes of footage, SAD-based detectors would require around seven minutes of footage to be reviewed and stored as evidence, in addition to the three minutes of footage of interest. This compares very poorly with the other techniques. It is quite surprising just how effective the use of average luminance is as a detection discriminator, as this technique can generate very low false alarm rates while maintaining a high rate of detection. For this reason, we have identified average luminance as the best candidate to form the basis of a motion detection metric. Enhancements to this technique might include improved tracking and also the implementation of block-based techniques to provide both improved discrimination and location estimation.

Table 1: Summary of Detector Performance

Detector	Detections	False Alarms	False Alarm Overhead
Entropy	9	18	1 min 12 sec (4%)
Average Luminance	9	1	4 sec (0.2%)
SAD	8 (1 miss)	101	6 min 44 sec (24%)
Maximum Macroblock SAD	9	117	7 min 48 sec (28%)

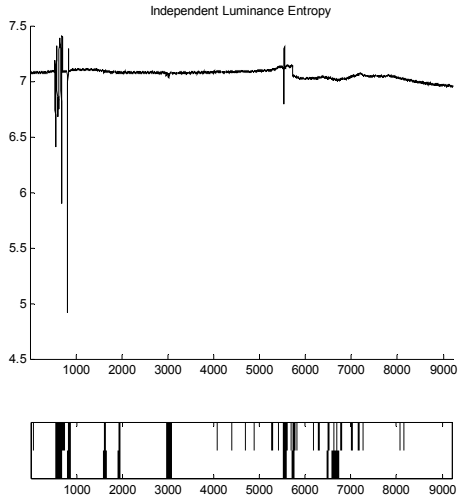

Figure 1: Performance of Independent Luminance Entropy. The bars at the bottom of the figure show the detected events (top) and the true motion events (bottom)

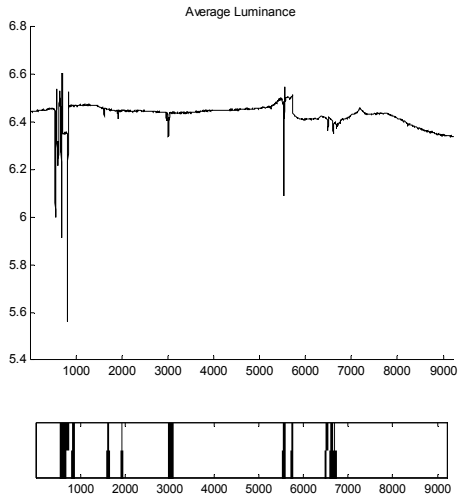

Figure 2: Performance of Average Luminance

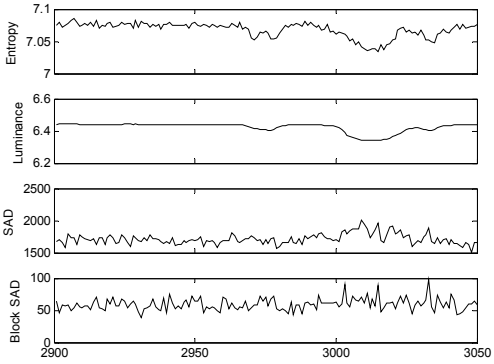

Figure 5: Magnified view of the detection metrics for the motion event at approximately frame 3000.

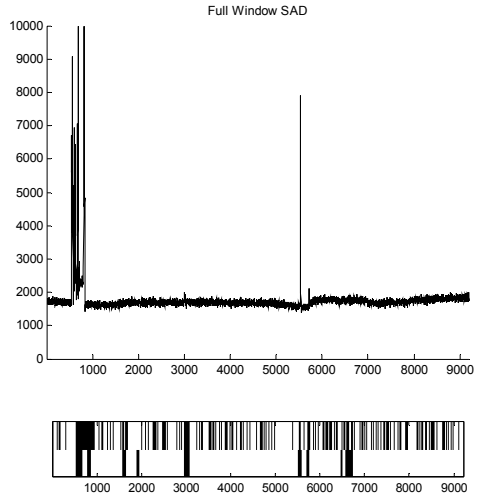

Figure 3: Performance of Full-Window Sum of Absolute Differences. Note the high false-alarm rate required to achieve acceptable detection.

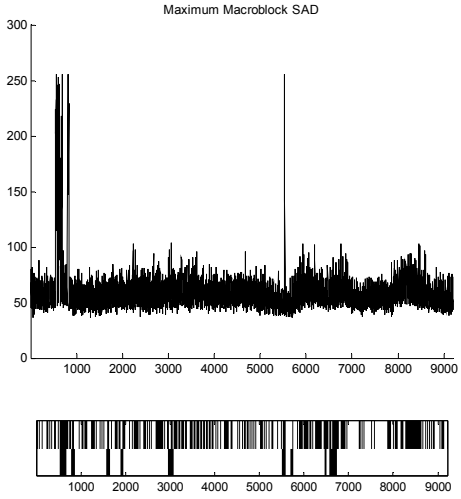

Figure 4: Performance of Maximum Macroblock Sum of Absolute Differences

Frame 3000 Frame 3009

Figure 6: At around frame 3000, a figure casts a shadow on frosted glass. Figure 5 shows that this event is barely detectable using Sum of Absolute Difference techniques, but is readily detected by average luminance and entropy.

7. LOCAL AVERAGED LUMINANCE

Based on the results given in Section 6, the effect of averaging blocks of 8x8 and 16x16 pixels was investigated. In effect this is Sum of Absolute Differences of average values over 64 and 256 pixels respectively, with the intention of improving performance by averaging, while providing localised motion statistics which are useful for further motion discrimination. The results are shown in Figure 7. It should be noted 16x16 averaging introduces a false alarm overhead of approximately 30 seconds, while 8x8 averaging introduces a false alarm overhead of approximately 1 minute 20 seconds.

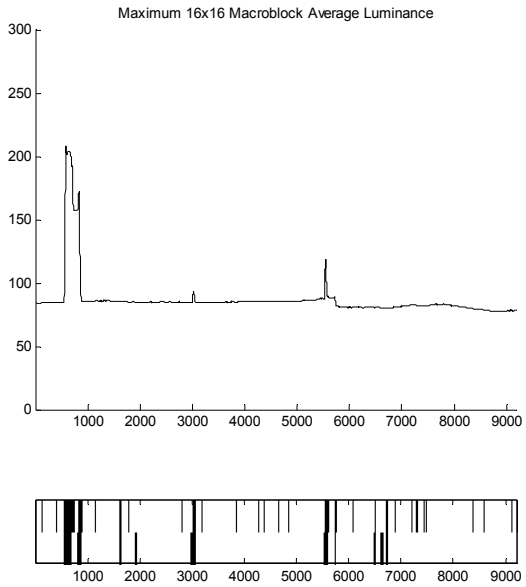

Figure 7: Performance of Maximum 16x16 Macroblock Average Luminance

8. CONCLUSIONS

We have presented an outline of the requirements for a system to automatically extract infrequent motion events from surveillance footage with a high level of reliability, in order to meet the burden of proof in court.

We have also proposed four candidate motion detection metrics and compared them using sample footage consisting of 9000 video frames taken over 30 minutes. Experimental results demonstrate that motion detection techniques based on Sum of Absolute Difference metrics are relatively ineffective, and that significantly better results can be achieved using metrics based on average luminance or signal entropy.

Further work is clearly required. Automatic setting of thresholds is highly desirable, especially with adaptation to light and activity levels, in a manner similar to the so-called Constant False-Alarm Rate (CFAR) radar algorithms described in [9, 11]. In addition the results need to be extended to capture metrics which are indicative of the type of motion event. These might include estimates of the size, position and velocity of moving objects, and some degree of event classification. It should also be clear that the outputs of our proposed system would be an effective foundation for the wide variety of proposed techniques for semantic interpretation of motion captured in video surveillance.

9. REFERENCES

[1] Al-Mualla, M.E, Canagarajah, C.N., and Bull, D.R. 2002., Video Coding for Mobile Applications, Academic Press.

[2] Cernekova, Z., Nikou, C., and Pitas, I., 2002. Entropy Metrics used for Video Summation, Proceedings of the 18th Spring Conference on Computer Graphics, 2002, pp 73-82. DOI = http://portal.acm.org/citation.cfm?id=584471&dl=

[3] Dee, H.M., and Velastin, S.A., 2007. How close are we to solving the problem of automated visual surveillance? A review of real-world surveillance, scientific progress and evaluative mechanisms, in Machine Vision and Applications, 5 May 2007. DOI = 10.1007/s00138-007-0077-z

[4] Duque, D., Santos, H., and Cortez, P., 2006. The OBSERVER: An Intelligent and Automated Video Surveillance System, in Lecture Notes in Computer Science, Springer, ISSN 0302-9743. 4141 (2006) pp 898-909.

[5] Hu, W., Tan, T., Wang, L., and Maybank, S., 2004. A Survey on Visual Surveillance of Object Motion and Behaviors, IEEE Trans. Systems, Man, and Cybernetics, Vol 34, No 3, August 2004

[6] Huang, Z.-K., and Liu, D.-H., 2007. Unsupervised Image Segmentation Using EM Algorithm by Histogram, in Lecture Notes in Computer Science, Springer, ISSN 0302-9743. DOI 10.1007/978-3-540-74171-8_130

[7] Konrad, J., 2000. Motion Detection and Estimation, in Handbook of Image & Video Processing, Bovik, A., (ed) ISBN 0-12-119790-5, pp 207-225

[8] Ngan, P.M., 1997. Motion Detection using Approximate Entropy, in Proceedings of the Digital Image Computing Techniques and Applications/Image and Vision Computing, Massey University, New Zealand, 1997, pp 379-384

[9] Poor, H.V., 1994. An Introduction to Signal Detection and Estimation, 2nd Ed, Springer-Verlag, ISBN 0-387-94173-8

[10] Sahoo, P.K., Slaaf, D.W., Albert, T.A., 1997. Threshold selection using a minimal histogram entropy difference, Optical Engineering, Vol 36, Issue 7, July 1997 pp 1976-1981. DOI = 10.1117/1.601404

[11] Van Trees, H.L., 1968. Detection, Estimation, and Modulation Theory, Part I – Detection, Estimation and Linear Modulation Theory, Wiley, ISBN 471-89955-0