Background Initialization with A New Robust Statistical Approach

Hanzi Wang and David Suter

Institute for Vision System Engineering
Department of. Electrical. and Computer Systems Engineering
Monash University, Clayton 3800, Victoria, Australia
{hanzi.wang; d.suter}@eng.monash.edu.au

Abstract

Initializing a background model requires robust statistical methods as the task should be robust against random occurrences of foreground objects, as well as against general image noise. The Median has been employed for the problem of background initialization. However, the Median has only a breakdown point of 50%. In this paper, we propose a new robust method which can tolerate more than 50% of noise and foreground pixels in the background initialization process. We compare our new method with five others and give quantitative evaluations on background initialization. Experiments show that the proposed method achieves very promising results in background initialization.

1. Introduction

There are many practical applications of tracking/surveillance: including monitoring freeways [1], recognizing human action [2, 3], motion segmentation [4, 5], etc. Effectively detecting and extracting moving foreground objects is a crucial step in these applications. To extract foreground objects, one usually needs to model the background scene using a short training video sequence. A number of background modeling methods have been proposed in recent years, e.g., [2, 3, 5, 6, 7, 8, 9]. However, most of these methods build up the background models, assuming that the training sequence is *free of foreground objects*. In many practical tasks, for example, in a busy road or in a public area, we must initialize the background model in a way that robust to the presence of foreground objects in the background training data. This problem, called *bootstrapping* [7], has received relatively little attention.

In this paper, we propose a new robust method for background initialization. The major advantage is that the proposed method can tolerate over 50% of noise in the data (including foreground pixels), in contrast with methods using the Median statistic which will break down totally when background constitutes less than 50% of the training data. A number of experiments are presented to show the advantages of the proposed method over other methods.

This paper is organized as follows: in section 2, we provide a short review of background modeling. We develop a robust method for background initialization in section 3. In section 4, experiments showing the advantages of our method are provided. We conclude in section 5.

2. Related work

Background modeling is mainly composed of three parts: model representation, model initialization, and model maintenance [4]. Much effort has concentrated on model representation and model maintenance. Early studies represent a background feature by an average of either grey-level or color samples at each pixel over a training time. To tolerate the influence of image noise, some statistical models are employed. One prominent example is Pfinder [3]. Pfinder assumes that the pixels, over a time window at a particular image location, are Gaussian distributed. After the background value of the pixel is obtained, exponential smoothing is employed to update for slow or gradual change in the background scene. Such approaches do not address scenes with dynamic backgrounds or where foreground objects are present in the training stage.

Many methods have been proposed for modeling dynamic background scenes. For example, Mixture of Gaussians (MOG) [6, 9, 10]. In MOG, the background features are characterized by a mixture of several Gaussians. Each Gaussian represents a distribution per pixel. Thus, MOG can efficiently model dynamic background scenes. However, when the background involves a wide distribution in color/intensity, modeling the background with a mixture of a small number of Gaussian distributions is not efficient. When foreground objects are included in the training frames, MOG will misclassify [7].

To improve MOG, a non-parametric method for background modeling was proposed [5]. However,

0-7803-9424-0/05/$20.00 ©2005 IEEE.

several pre-calculated lookup tables for the kernel function values are required to reduce the burden of computation of this approach. Also, this method can not resist the influence of foreground objects in the training stage.

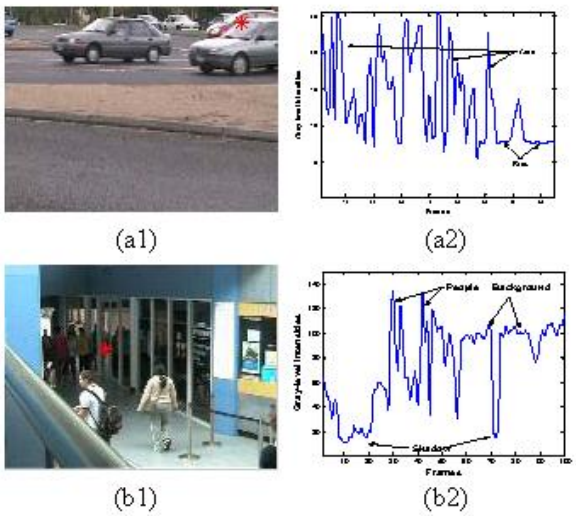
In contrast to background model representation and model maintenance, only a few studies of background model initialization have been made (e.g., [1, 4, 11, 12]). In [12], a Smoothness Detector (SD) Method was proposed. They assumed that a background value always has the longest stable value. At each pixel, a moving window along time is employed to search for the stable intervals. However, we find one problem of the method is that when the data include multi-modal distributions (i.e., some modes from foreground objects and some modes from background as shown in Figure 2 and Figure 3), and when the modes from foreground objects tend to be relatively stable, this method can not differentiate these modes from those from the background.

In order to decide the window length L and the intensity flicker of the window T_f for each pixel, [12] proposed an Adaptive Smoothness Detector (ASD) method. Because the ASD method tries different L and T_f at each iteration until the solution is found, the computational cost of the ASD method is high.

Motivated by [12], a Local Image Flow (LIF) algorithm [11] was proposed. Two steps are used: in the first step, all stable sub-intervals in a training sequence are located for each pixel. In the second step, the method locates the sub-interval with the greatest average likelihood using local motion information, and produces a background value by computing the mean value over the chosen sub-interval. Optical flow is computed for each consecutive pair of images and used to estimate the likelihood. While this potentially adds valuable information, most optical flow computation methods themselves are computationally complex and very sensitive to noise.

In [1], the authors used the median intensity value over observations at each pixel, to initialize the background for a traffic monitoring system. The underlying assumption is that the background at each pixel can be seen for more than 50 percent of time in the training sequence. However, the requirement that background appear more than 50% of time in a video sequence may not be always satisfied. Figure 1 illustrates two such examples. In Figure 1, we can see that the background value at the marked pixel (with red star) is visible less than 50 percent of the training time. The noise is either from the moving foreground objects or the shadows of the foreground objects.

A robust method which can tolerate more than 50% of noise is possible [13]. Examples include RANdom Sample Consensus (RANSAC) [14], Adaptive-Scale Sample Consensus (ASSC) [15], etc. To overcome the problems inherent in methods based on the Median, we introduce a consensus-based robust method of background initialization. The details of the proposed method will be introduced in the next section.



(a1) (a2)

(b1) (b2)

Figure 1: Two examples that background is visible less than 50 percent of the training time: (a1) and (b1) show one frame of each training sequence; (a2) and (b2) show the intensity distributions over time at one pixel (marked by red star) of the sequence.

3. The proposed method for background initialization

3.1 Assumptions

Our assumptions are similar to those in [11, 12]:

1. The background at each pixel should be revealed at least for a short interval during the training period.

2. A background value tends to be relatively stable and constant.

3. A foreground object can remain stationary for a short interval in the training sequence. However, the interval should be no longer than the interval from the revealed static background (in assumption 1).

4. The background scene remains relatively stable.

Stability is one characteristic of essentially stationary backgrounds. The foreground value at a pixel is assumed to have no less variance in grey-level intensity than a background value.

When the background involves dynamic scene (such as waving trees, rain, etc) the second and the third assumptions are invalid. To the best of our knowledge, all the proposed background initialization

methods have concentrated on handling stationary backgrounds.

3.2 The Proposed method

Our method employs a two-step framework: (1) locate all non-overlapping stable subsequences of pixel values; (2) choose the most reliable subsequence, and use the mean value of either the grey-level intensities or the color intensities over that subsequence as the model background value.

In the first step, we use a sliding window with a minimum length L_w, similar to the work in [11, 12], to locate all stable sub-intervals $\{l_k\}$. For a test sequence of N frames, we have N observations at each pixel $\{x_i \mid i = 1,...,N\}$. Let $x_{l_k(t)}$ be a pixel value of the kth subsequence l_k at time t. The kth stable subsequence candidate should satisfy the following equation:

$$\forall (t-1,t) \in l_k, \begin{cases} \left| x_{l_k(t)} - x_{l_k(t-1)} \right| \le T_f \\ \left| x_{l_k(t)} - \bar{x}_{l_k(t-1)} \right| \le T_f \end{cases} \quad (1)$$

where $\bar{x}_{l_k(t-1)}$ is the mean value over an interval from the beginning of the subsequence l_k to time t-1.

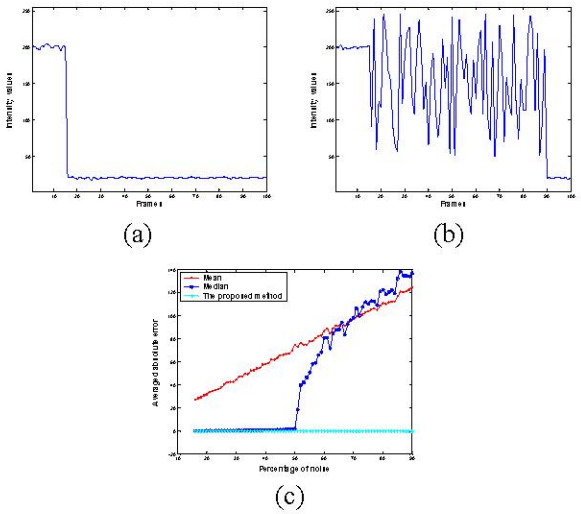
If we can not find any subsequence candidate with a minimum length L_w along time, we use the longest stable subsequence from the candidates. In our experiments, we experimentally set L_w to 5 and T_f to 10, for all test sequences. The chosen subsequences can contain pixels from foreground, background, shadows, highlights, etc. (e.g., see Figure 1 b). We need to further process these subsequences in the next step.

The second step is a crucial step, because in this step, a reliable subsequence which is most likely to arise from the background will be chosen. The authors in [11] used local motion information (optical flow) for choosing the reliable subsequence. However, optical flow methods are computationally expensive and they suffer many problems: such as aperture, sensitivity to noise (e.g., shadows, illumination changes), etc.

Our definition of reliability is motivated by RANSAC [14] and other robust methods. We build in to our objective function the notions of consensus and of scale estimation. We consider both the number (n) of data points "agreeing" with a model (contained in the candidate interval), and the distribution of these data (e.g. standard variance S), in our objective function: n should be large, and S should be small. We therefore define our objective function as finding the most stable interval from the non-overlapping sub-intervals $\{l_k\}$ by:

$$\hat{l}_k = \arg\max_k (n_{l_k} / S_{l_k}) \quad (2)$$

where n_{l_k} and S_{l_k} are respectively the number of values (length of) and the standard variance of the observations in the kth subsequence l_k.



(a) (b) (c)

Figure 2: Estimating background value from noisy data: (a) and (b) illustrate two cases of the distributions of the simulated data; (c) the results obtained by the three methods.

To illustrate the robustness of the proposed method we generate synthetic data to simulate the observations over time at a pixel. One hundred data values (i.e., 100 frames) were generated. The first fifteen data values (i.e., a relatively stationary foreground object pixel) have intensity value of 200 and standard variance of 2. From the sixteenth to the i'th data, we simulate random noise (such as foreground objects in transit at that pixel) with intensity values ranging from 50 to 250. We simulate a background value in the sub-interval from the (i+1)'th data to the 100th data, with unit variance. We increase i from 16 to 90 with step 1 each time. We repeat the experiment ten times and output the average value as results.

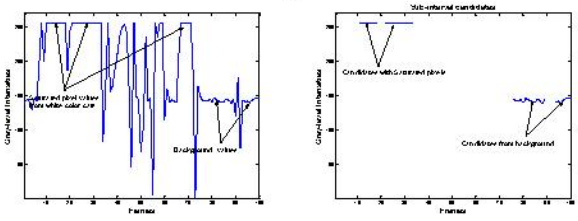
Two simulated data distributions with i=16 and i=90, are shown in Figure 2 (a) and (b). Figure 2 (c) shows the results by three statistics: Mean, Median, and the proposed method. From the results, the Mean is not robust to noise at all. The error by the mean is largely affected by the percentage of noise in the data and the distributions of the noise. Although the Median can tolerate the influence of noise, when the noise occupies less than 50 percent of the data, the Median

method breaks down. In contrast, the proposed method is much more robust to the influence of noise than the Median method.

Although the proposed method can achieve accurate results in most cases, we note that equation (2) might be erroneous when S_{l_k} is very small or zero. This can happen when some pixels of a short subinterval have saturated colors. The saturated pixel values are clipped within the range from 0 to 255 and sequences containing these saturated pixels have a very small (or zero) standard variance [16]. For this case, the assumption (2) in subsection 3.1 is violated. When we detect such a case happens, we use the following equation instead of equation (2).:

$$\bar{I}_k = \arg\max_k (n_{l_k}) \tag{3}$$

(a)



(b) (c)

Figure 3: One example showing that the intensities of the saturated pixels are clipped: (a) shows one frame of the video sequence (the pixel marked by a red star is investigated); (b) shows the intensity distributions over time at that pixel; (c) the possible sub-interval candidates.

Figure 3 shows an example where the intensities of some saturated pixels are clipped. Figure 3 (a) shows one frame of the test sequence. We investigate the grey-level intensity distribution of the observations at one pixel which was marked with a red colored star. In Figure 3 (b), we can see that there are some saturated pixels corresponding to white colored cars passing by. The sub-interval candidates obtained in the first step are shown in Figure 3 (c). The two candidates corresponding to saturated pixels have a standard variance of zero. In such case, we should use equation (3) instead of equation (2).

4. Experiments

The test sequences are recorded by a Canon MV750i digital video camera. We stored the sequences at a resolution of 160x120, and a sample rate of five frames per second. We have deliberately chosen different background including both indoors and outdoors scenes, including foreground objects, shadows, highlights, and illumination changes to simulate true situations that a visual surveillance system may meet in practice.

Road1 (**R1**): Heavy traffic in daytime (some shadows on the road).

Road2 (**R2**): Vehicles passed by a crossing road in the evening. Some parts of the road were highlighted when vehicles (with lights on) got close to those parts.

Train Station (**TS**): A gate of a train station. Many people exited or entered the station through that gate.

Sport Center (**SC**): In an indoor sport center, people walked through a corridor. Shadows of people were cast on the glass wall and the floor of the corridor. Also some illumination changes happened when people exited the back door and covered the light outside.

Pharmore Shop (**PS**): A pharmacy shop, which is located inside a big shopping center. People walked in front of the shop. The illumination of the background scene sometimes changed because of the reflected sunlight outside the shopping center.

We compare the proposed method with five other methods. All of the methods perform at pixel-level for background initialization (methods based on area can be expected to achieve better results at great cost). To test each method, we choose two sub-sequences (S1 and S2) which include a number of frames ranging from 30 to 100 in each sub-sequence, from each test sequence. To evaluate the performance of each method, we employ three criteria, similar to those used in [11]: a) the Average gray-level Error (AE); b) the Number of Error pixels (NE); and c) the Number of Clustered error pixels (NC). We use the Mean value of Total error (MT) of the ten sub-sequences over each criterion as the overall measurement for each method.

We generate a Reference Frame (RF) for each test sequence by using the mean value of selected frames that are free of foreground objects. An error pixel is one whose grey-level value differs from the value of the reference pixel by a threshold 20. We define a clustered error pixel when the 4-connected neighbors of that error pixel consist of more than 4 error pixels.

Figure 4 shows one frame of each test subsequences and the resulting error pixels (corresponding to the white color pixels), obtained by the five other methods

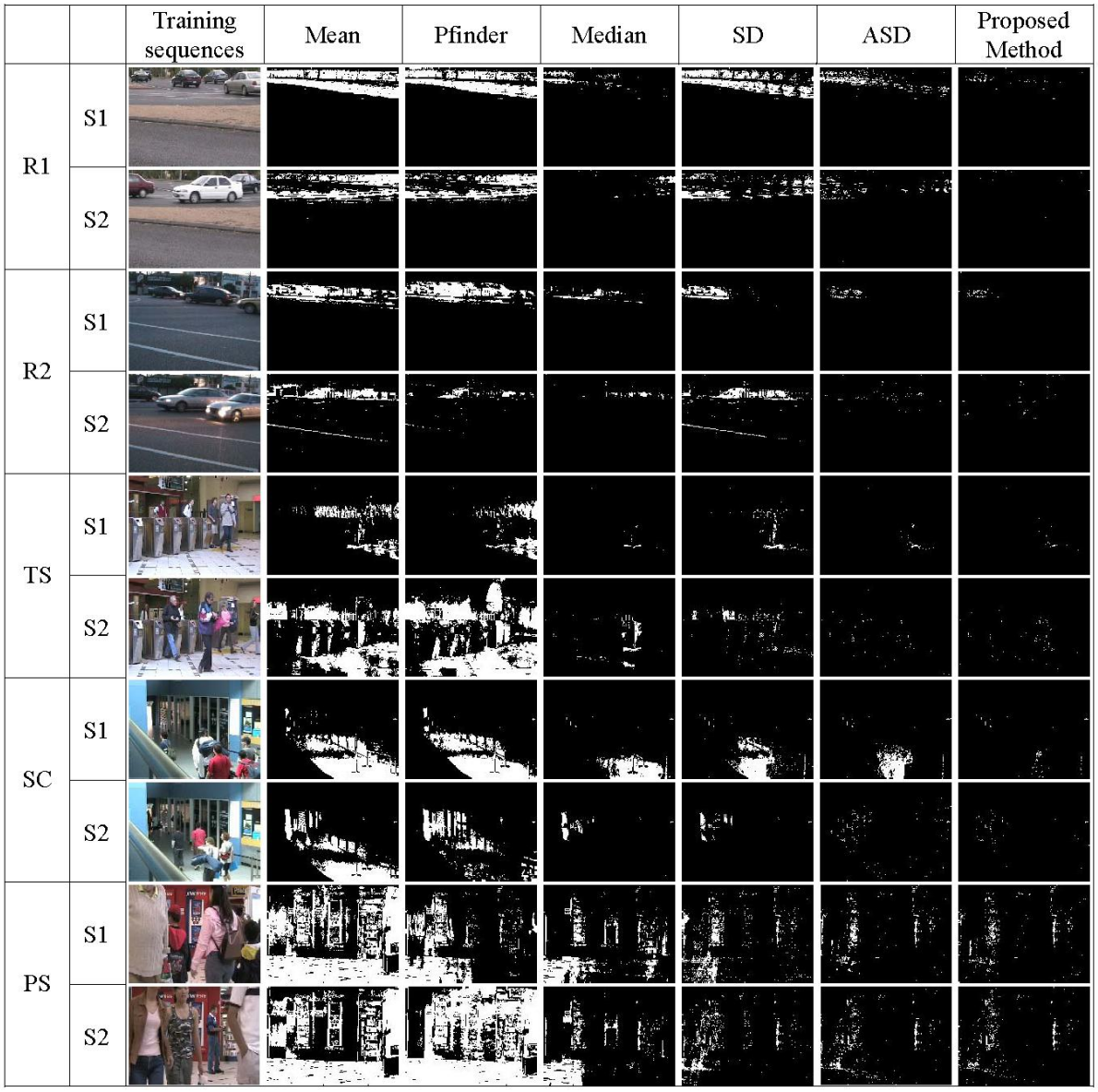


Figure 4: Ten video sub-sequences of five test videos. The third column shows one frame of each training subsequence; the remaining columns show the difference between the background and the background estimate obtained by the competing methods. The results obtained by the proposed method are shown in the last column.

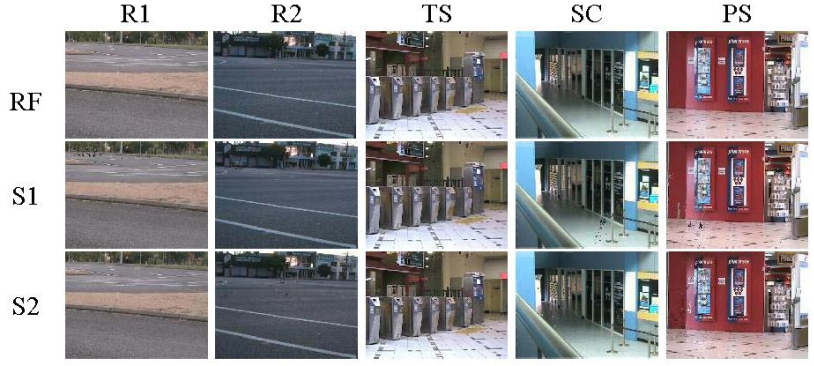


Figure 5: The reference background and the initialized background images by the proposed method.

		R1		R2		TS		SC		PS		MT
		S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	
Mean	AE	9.61	10.27	5.79	9.10	5.81	14.12	11.69	8.75	26.48	25.12	12.67
	NE	2994	2369	1630	1320	1323	4992	3537	3102	10253	9799	4132
	NC	2965	2273	1571	1231	1211	4811	3436	3023	10031	9677	4023
Pfinder	AE	9.14	9.17	6.25	6.01	5.50	20.89	10.08	9.92	12.99	38.82	12.88
	NE	2790	2127	1917	411	1125	7805	3402	1822	3969	12690	3806
	NC	2752	2016	1866	312	1042	7605	3347	1699	3746	12573	3696
Median	AE	5.14	4.58	2.69	3.45	2.89	4.15	6.63	3.14	9.51	8.49	5.07
	NE	352	159	276	142	40	353	1349	271	2559	2092	759
	NC	282	127	239	114	28	296	1301	247	2347	1947	693
SD	AE	7.99	5.94	2.83	5.58	2.96	3.50	6.10	2.77	7.85	5.43	5.10
	NE	2097	976	515	872	226	399	1304	217	1400	921	893
	NC	2018	840	487	741	153	228	1195	181	961	603	741
ASD	AE	5.59	6.01	2.43	3.58	2.66	2.81	7.62	2.94	6.47	4.64	4.48
	NE	588	252	114	55	44	56	892	123	598	559	328
	NC	443	152	82	11	22	0	819	15	420	306	227
The proposed method	AE	4.33	4.32	2.05	3.00	2.54	2.77	2.81	2.46	6.27	4.36	3.49
	NE	70	10	57	37	21	63	76	51	541	484	141
	NC	23	0	23	4	7	5	28	15	296	238	64

Table 1: Experimental results by different methods on test sequences.

and the proposed method. A quantitative comparison is given in Table 1. From these results, we can see that the Mean and the Pfinder methods are the most inaccurate in background initialization. The Mean takes all observations at each pixel in the test subsequence into account. The Pfinder, using a temporal smoothing technique, gives larger weight value to recent observations. When the observations contain pixels from other than background, these two methods break down.

Compared with the Mean and the Pfinder, the Median method achieves a much better result because of its robustness to noise (from foreground objects, shadows, etc.). However, when the test subsequence includes too many foreground objects, or if the background value is visible for less than 50 percent of the test subsequence (more noticeable, in the S1 of Sport Center sequence, and in the S1 and S2 of the Pharmore Shop sequence), the Median method fails to estimate the background.

SD obtained more accurate results than the Median in the SC and PS sequences, but less accurate results in the R1, R2, and TS sequences. ASD achieves better results than the SD method in all test sequences because it uses different window length L and T_f at each pixel location. However, the cost is about 30-50 times slower than SD in computational time.

Among the six methods, the proposed method achieves the most accurate results and it also is about three times faster than SD, and about 100 times faster than ASD.

We show the initialized backgrounds for the test sequences in Figure 5. We use the mean values of the RGB colors of the chosen sub-interval as the initialized background. The reference images of the test video sequences are shown in the second row. The initialized background scenes by using two subsequences (S1 and S2) of each test video are respectively shown in the third and the fourth rows. The proposed method obtains good results in background initialization for most of the test sequences. However, for the PS sequence, the results include relatively more error pixels. Most of the error pixels are caused by the illumination changes.

5. Conclusion

In this paper, we develop a new robust method for background initialization. The proposed method can be used in many places where foreground objects can not be avoided. The main strength of the proposed method is in that its high robustness to noise in data and the method is a great improvement over the traditional Median method.

We have evaluated our method with several other methods on various outdoor and indoor video sequences. Experimental results on background initialization have shown that our method outperforms

other methods and can achieve very promising results even when background is revealed much less than half of time in the training sequences.

Acknowledgements:

This work is supported by the Australia Research Council (ARC) grant DP0452416. The work was carried out within the Monash University Institute for Vision Systems Engineering.

6. References

1. B. Gloyer, et al. "*Video-based Freeway Monitoring System Using Recursive Vehicle Tracking,*" in *Proc. of IS&T-SPIE Symposium on Electronic Imaging: Image and Video Processing*, 1995.
2. I. Haritaoglu, D. Harwood and L.S. Davis, "*W4: Real-Time Surveillance of People and Their Activities,*" PAMI. **22**(8): p. 809-830, 2000.
3. C.R. Wren, et al., "*Pfinder: real-time tracking of the human body,*" PAMI. **19**(7): p. 780-785, 1997.
4. M. Cristani, M. Bicego and V. Murino. "*Multi-level background initialization using Hidden Markov Models,*" in *First ACM SIGMM international workshop on Video surveillance*: p. 11 - 20, 2003.
5. A. Elgammal, et al., "*Background and Foreground Modeling using Non-parametric Kernel Density Estimation for Visual Surveillance,*" Proceedings of the IEEE. **90**(7): p. 1151-1163, 2002.
6. C. Stauffer and W.E.L. Grimson. "*Adaptive Background Mixture Models for Real-time Tracking,*" CVPR: p. 246-252, 1999.
7. K. Toyama, et al. "*Wallflower: Principles and Practice of Background Maintenance,*" ICCV: p. 255-261, 1999.
8. M. Harville. "*A Framework for High-Level Feedback to Adaptive, Per-Pixel, Mixture-of-Gaussian Background Models,*" ECCV: p. 543-560, 2002.
9. N. Friedman and S. Russell. "*Image Segmentation in Video Sequences: A Probabilistic Approach,*" in *Proc. Thirteenth Conf. on Uncertainty in Artificial Intelligence*: p. 175-181, 1997.
10. M. Harville, G. Gordon and J. Woodfill. "*Foreground Segmentation Using Adaptive Mixture Models in Color and Depth,*" in *IEEE Workshop on Detection and Recognition of Events in Video*: p. 3-11, 2001.
11. D. Gutchess, et al. "*A Background Model Initialization Algorithm for Video Surveillance,*" ICCV: p. 733-740, 2001.
12. W. Long and Y.H. Yang, "*Stationary Background Generation: An Alternative to the Difference of Two Images,*" PR. **23**(12): p. 1351-1359, 1990.
13. C.V. Stewart, "*Robust Parameter Estimation in Computer Vision,*" SIAM Review. **41**(3): p. 513-537, 1999.
14. M.A. Fischler and R.C. Bolles. "*Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,*" Commun. ACM. **24**(6): p. 381-395, 1981.
15. H. Wang and D. Suter. "*Robust Adaptive-Scale Parametric Model Estimation for Computer Vision,*" IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 26, No.11, pages 1459-1474, 2004.
16. T. Horprasert, D. Harwood and L.S. Davis. "*A Statistical Approach for Real-Time Robust Background Subtraction and Shadow Detection,*" in *ICCV'99 Frame-Rate Workshop*, 1999.