Multiple Mobile Objects Detection and Tracking with an Overhead Camera

Andrzej Kasinski and Alaa Hamdy

Institute of Control and Computer Engineering, Poznan University of Technology, ul. Piotrowo 3A, 60-965 Poznan, Poland, Email: akas,hamdy@ar-kari.put.poznan.pl

Abstract

This paper presents the results of using a number of proprietary algorithms aimed at the real-time moving objects detection. The objects are mobile robots and human beings operating on a complex and unevenly illuminated background (textured floor). The motion detection algorithm is based on an adaptive background method that models each sub-sample pixel as a mixture of Gaussian distributions. This approach yields a stable detector dealing well with long-term scene changes, however it involves massive computations, and moreover it takes fast variations of the scene illumination as a motion evidence in a sequence of images. Thus, as a preprocessor to the motion detection algorithm, a local homomorphic filter is used to suppress the illumination component, which is time-variable, rendering the motion detection less sensitive to that environmental disturbance. As it adds an extra processing cost, we propose to use the combination of homomorphic filter with motion detector, which is restricted only to a subset of pixels (sampled on uniform grid). From the sub-sampled pixels, only the pixels belonging to the moving objects are selected as seed points for the region-growing processes to extract silhouettes of mobile objects. This is equivalent to the fusion of the object-attribute related information with the motion related information. The algorithm can run in real-time on a typical computer (Pentium II PC or higher).

1 Introduction

Automatic detection of moving objects in image sequences is difficult while the scene (background) is complex and textured. This is true especially, whenever a CCD camera is used and scene-illumination conditions are poor (weak exposition). Shadows and reflections additionally disturb the motion-based detection and mobile objects tracking.

Normally, in such circumstances, the detection and localization of mobile robots can be achieved by placing well detectable color light sources (beacons) in the corners of a platform [3]. In some other solutions, robots are equipped with infrared beacons. These beacons can be used to detect moving robots. However, in some cases, it is not possible to improve the visibility of moving objects (their passivity is required). So efficient methods, suitable for poor operating conditions (uneven and weak illumination, textured background) are needed. The goal of the article is to present a solution for the case when robots and other vehicles are not especially prepared but enter the observation space "as they are". In other words, we look for a real-time, image-based detector of passive mobile objects (robots and human beings) instead of beacon-based detection methods.

Many researchers working on the image-based detection have abandoned non-adaptive methods of backgrounding, as they usually involve manual initialization. Without re-initialization, errors in the background reconstruction would accumulate over time, making such methods useful only for highly supervised short-term tracking applications, where there is no significant changes in the scene.

A standard approach in adaptive backgrounding is to average the images over time [9], [4] and thus creating a background approximation which is similar to the current static scene, except for regions where the motion occurs. While this is efficient in situations where objects move continuously and the background is visible during a significant amount of the time, it is not robust in the case of scenes with many moving objects, especially when moving slowly. With such an approach, one cannot handle bimodal backgrounds, the uncovered background is recovered slowly, and the single threshold has to be used (predetermined for the entire scene).

The paper is organized as follows: Section 2 describes how to deal with time-variable illumination. Section 3 discusses the motion detection by adaptive background subtraction. Section 4 describes how to use the seeded-region growing to extract the silhouettes of mobile objects. Conclusions and final remarks are given in section 5.

2 Dealing with time-variable illumination

The image is typically formed by recording the light reflected from an object that has been illuminated by some external light-sources. Based on this observation, one simple model of the image is the product of the illumination component **i** and the reflectance component **r**. The illumination component is usually varying slowly, while the reflectance component is assumed to vary rapidly in time [8].

In order to split the image function components, a logarithmic operation is applied to pixels of the Region of Interest (ROI), transforming the multiplicative image-function factors to the additive ones. The result is then low-pass filtered to obtain log (**i**), and high-pass filtered to obtain log (**r**). Once the last two components have been extracted, log (**i**) is attenuated and log (**r**) is emphasized to increase the local contrast. The attenuated log (**i**) component and the emphasized log (**r**) component are then combined and the result is exponentiated to get back to the image intensity domain as it is shown in figure 1.

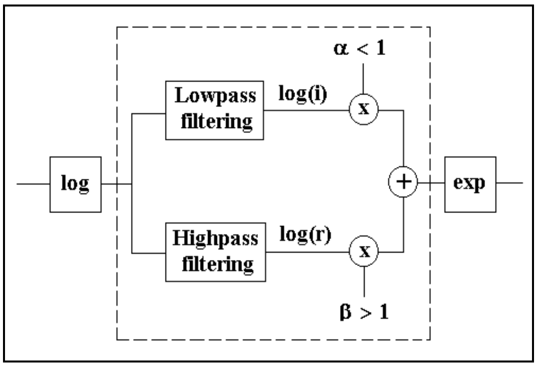


Fig. 1. Block diagram of the homomorphic filter.

In our case, the whole image is uniformly subsampled by selecting pixels over a rectangular sampling grid and only the subsampled pixels of the image are homomorphically filtered. Low-pass filter used is a Gaussian-kernel, whose output is subtracted from the logarithmic original, yielding a high-pass component [2]. Exponentiation of high-pass component brings the reflectance component almost separated. The lowpass kernel of size 51 x 51 has been used. To speed up the filtering, the Gaussian separability is used [12]. This means that we use one-dimensional Gaussian function of size 51. The input image is convolved first with a vertical one-dimensional Gaussian. Its output is used as the input to a horizontal one-dimensional Gaussian convolver. Moreover, lookup tables have been implemented to speed up filtering computation [13]. Namely, we used lookup tables to perform the logarithm operation and the filtering computation as well, arriving to the multiplication-free algorithm. Figure 2 demonstrates the effect of homomorphic filtering. It is clearly visible that in the reflectance image (**r**) illumination effects have been strongly suppressed, while object information has been preserved. This effect is particularly strong while looking at the image-sequence registered under dynamic illumination.

From the practical point of view, the separation works well enough. Of course, the reflectance image still contains low-frequency residuals from the illumination, as the separation of the two components is not perfect.

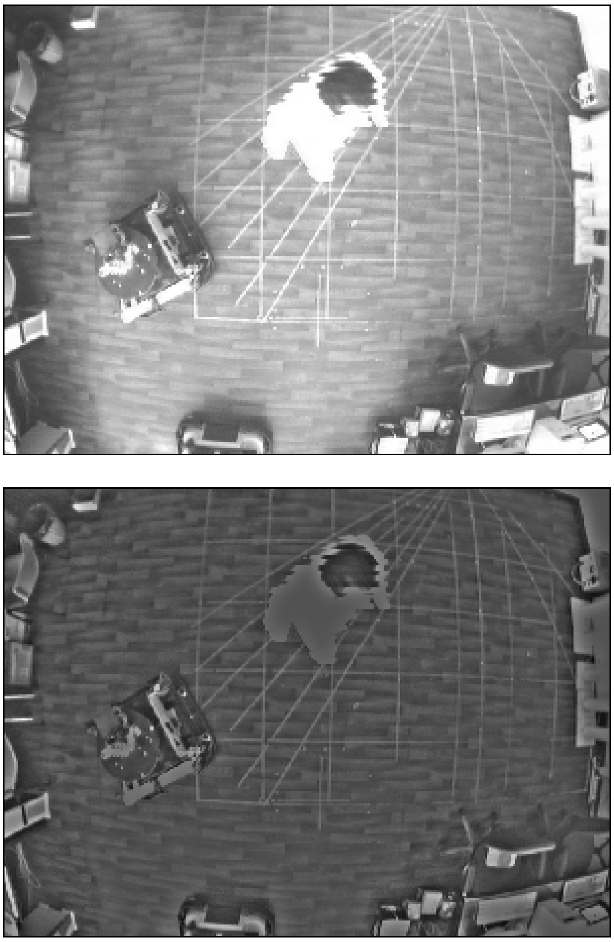


Fig. 2. The effect of homomorphic filtering on a particular image from the sequence, without filtering (top), with filtering (bottom).

3 The adaptive subtraction of the background

In order to reduce the computational cost of the reflectance component extraction, the homomorphic filter has been applied only to sub-sampled pixels and the result is merely an approximation of the reflectance distribution. At this point, the adaptive-background method for motion detection [1], [10], [11] is applied only w.r.t. these extracted sub-sampled pixels, i.e., one is looking to distinguish the foreground/background only in these locations of pixels. To achieve it, each pixel of the sampling grid is modeled as a mixture of five Gaussian distributions (for each sample-pixel five mean values, five variance values and five weight values at time t are maintained). At any time t, what is known about a particular pixel $\{x_0, y_0\}$, is its history:

$$\{X_1,...,X_t\} = \{I(x_0, y_0, i) : 1 \le i \le t\} \tag{1}$$

where I is the image sequence.

The probability of relating the current pixel value with an appropriate model is:

$$P(X_t) = \sum_{i=1}^{K} w_{i,t}\, \eta\left(X_t, \mu_{i,t}, \Sigma_{i,t}\right) \tag{2}$$

where:
K - is the number of Gaussian distributions used,
$\omega_{i,t}$ - is an estimate of the weight of the ith Gaussian in the mixture at time t,
$\mu_{i,t}$ - is the mean value and $\Sigma_{i,t}$ - is the covariance matrix of the i^{th} Gaussian in the mixture at time t,
η - is a Gaussian probability density function.

$$\eta\left(X_t, \mu, \Sigma\right) = \frac{1}{\left(2\pi\right)^{\frac{n}{2}} \left|\Sigma\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t-\mu_t)^T \Sigma^{-1} (X_t-\mu_t)} \tag{3}$$

Each new extracted pixel value of the sub-sample is checked against the existing five Gaussian distributions until the match is found. The matching criterion is defined as a pixel's value lying within 2.5 standard deviations from the mean of a current distribution. The mean and variance parameters of the distribution, which matches such a new observation, are updated as in [1]:

$$\mu_t = \left(1-\rho\right)\mu_{t-1} + \rho\left(X_t\right) \tag{4}$$

$$\sigma_t^2 = \left(1-\rho\right)\sigma_{t-1}^2 + \rho\left(X_t - \mu_t\right)^T \left(X_t - \mu_t\right) \tag{5}$$

where ρ is the learning dynamics coefficient.

$$\rho = \alpha\eta\left(X_t \middle| \mu_k, \sigma_k\right) \tag{6}$$

and α is the learning rate coefficient (forgetting factor).

The prior weights of the distribution at time t are adjusted as follows:

$$\omega_{k,t} = \left(1-\alpha\right)\omega_{k,t-1} + \alpha\left(M_{k,t}\right) \tag{7}$$

where $M_{k,t}$ is 1 for the model which matched and 0 for the unmatched models (Gaussians).

If none of the five distributions matches the current reflectance pixel value, the least probable distribution is replaced with a distribution having an initial high variance, and a low *a priori* weight and the current reflectance value as its mean value. The mean and variance parameters for unmatched distributions are kept unmodified.

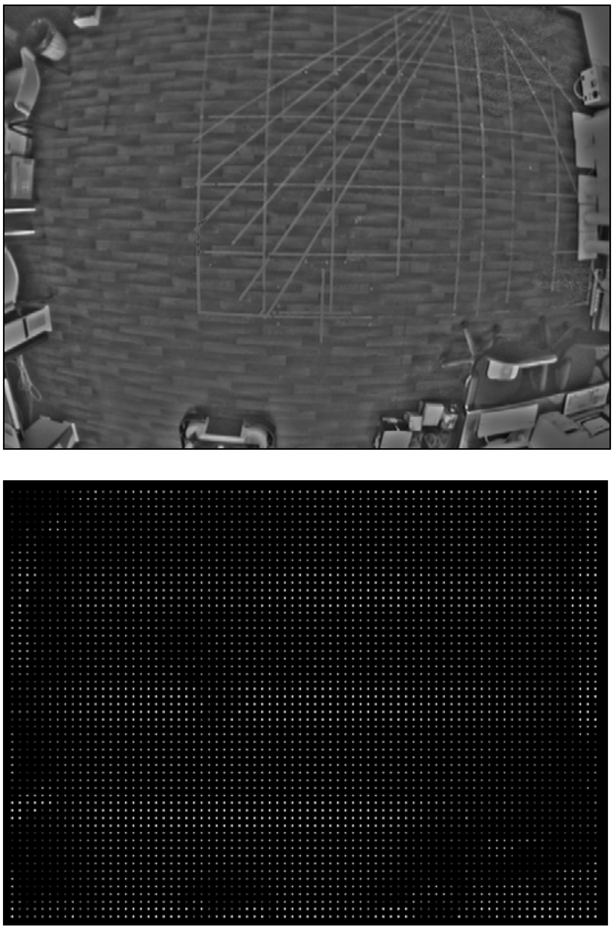


Fig. 3. The most probable background image without sub-sampling (top) and with sub-sampling (bottom).

The next step is to decide whether the sub-sample pixel belongs to the background. In order to do that, we sort all the components in the mixture in the order of decreasing ratio (ω/σ). In effect, this ratio assigns higher importance to those mixture components that have been supported by strongest evidence and had the lowest variance. The intuitive sense of this ratio is related to the fact, that the components, which correspond to the background, typically have more observations attributed to them and thus vary a little. Then, after the components are sorted, it is possible to set a threshold T, separating components representing the background pixels from the ones representing the foreground. The first B components of the sorted mixture are treated as related to the background. Now, if the pixel fits best one of the background models, it is marked as belonging to the background.

$$B = \arg\min_b \left(\sum_{k=1}^{b} \omega_k \succ T \right) \tag{8}$$

Figure 3 shows the most probable background image, displaying chosen dominant Gaussian mean for each pixel's mixture model both for the whole image and for the sub-sampled image. Pixel values that do not fit the background distributions are considered to belong to the foreground. This holds until a Gaussian including them is obtained, converting them to a new background mixture (which means that sufficient and consistent evidence supporting such a conversion is available).

It should be noted that the adaptive background subtraction algorithm is applied only on the subsampled pixels. The resulted sampled background is shown in figure 3 (bottom). Figure 3 (top) is given here only for comparison, as figure 3 (bottom) alone could be unclear.

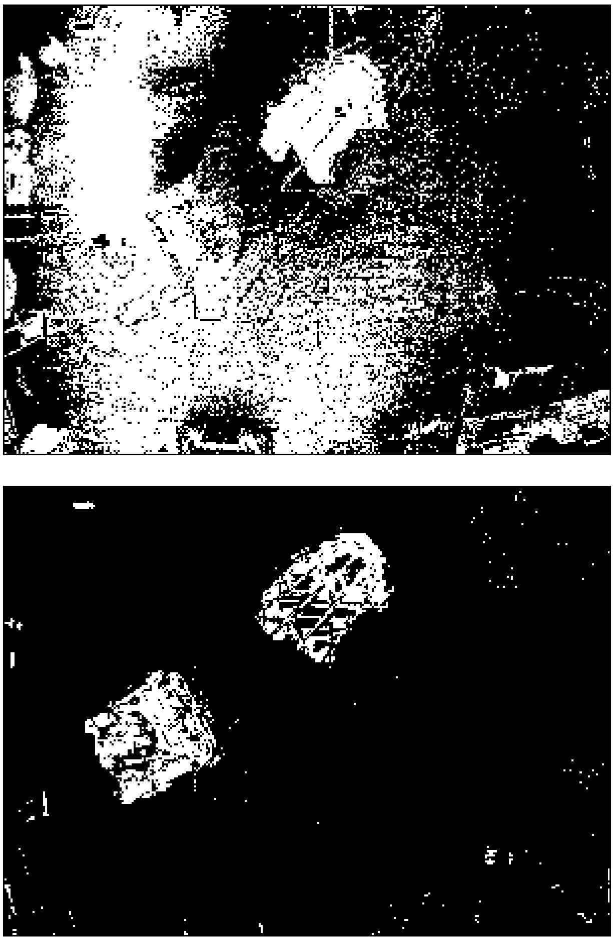


Fig. 4. The foreground pixels without preprocessing filter (top) and with preprocessing filter (bottom).

Variations in the illumination of the scene can cause problems for many backgrounding methods. In [1], one can deal robustly with normal lighting changes, but cannot deal with quick changes as caused by cloud cover for instance. These changes can sometimes cause the necessity of a new set of background distributions, which takes 10-20 seconds for the system to stabilize. Afterward, tracking will continue unhindered. Figure 4 (top) shows the foreground pixels in case of method [1] applied on the original frame sequence during quick change in illumination, the reference image is in Figure 2 (top). It is evident from figure 4 (top) that the obtained foreground pixels are not only due to apparent motion, but also are due to quick lighting changes. It is enough to compare this figure with figure 2 (top).

Figure 4 (bottom) displays the obtained foreground pixels in the case of method [1] as applied on the preprocessed frame sequence, during the same quick changes in illumination. The corresponding reference is here figure 2 (bottom). It is evident form figure 4 (bottom) that most of the foreground pixels are only due to apparent motion and not to quick lighting changes as before. It is enough to compare this figure with figure 2 (bottom). It should be noted too that, in figure 4 complete (non sub-sampled) images have been included only to enable the localisation of results. The sub-sampled foreground pixels exactly are displayed in figure 5 (top).

4 The Seeded-Region Growing and its robustness to the image-scale change

At this point, only the sub-sampled foreground pixels are obtained. From these pixels, we have to complete the missing foreground pixels in the frame. In other words, we have to extract the silhouettes of mobile objects as if the adaptive background subtraction algorithm would have been applied for the complete image. In order to do the job, the appropriate region-based segmentation method can be useful. One possible choice is a Seeded-Region Growing method (SRG).

In that case, each one of the sub-sampled foreground pixels is used as a seed-point and the enhanced version of the SRG described in [5], [6] is applied. From these seed-points region-growing processes start. The use of foreground pixels as seed points allows combining the motion information with the pixel-attribute information. A variable local threshold value used as stopping criterion for the pixel aggregation process is statistically estimated. A suitable area (a proper size window centered on the seed point) surrounding the initial seed point, as its center, is taken for calculation of the local mean μ and the local variance σ^2. The two values calculated as $\mu \pm \sigma$ are then used as thresholds similarly to [7]. Figure 5 (bottom) shows the results of segmentation with SRG method. The aggregated pixels are grouped into regions (blobs) by a two-pass connected components algorithm. The insignificant components (including few pixels) are then removed, resulting in erasing all artifacts from figure 5 (bottom).

Working on the extensions of the SRG method, leading towards its robustness w.r.t. the seed-point choice, it has been experimentally established that the SRG segmentation is moreover insensitive to the image-scale change (up to the spatial discretisation noise). Figure 6 demonstrates the stability of the segmentation results despite quite abrupt image-scale changes, for a single image-frame and one of mobile robots used in the experimental setup.

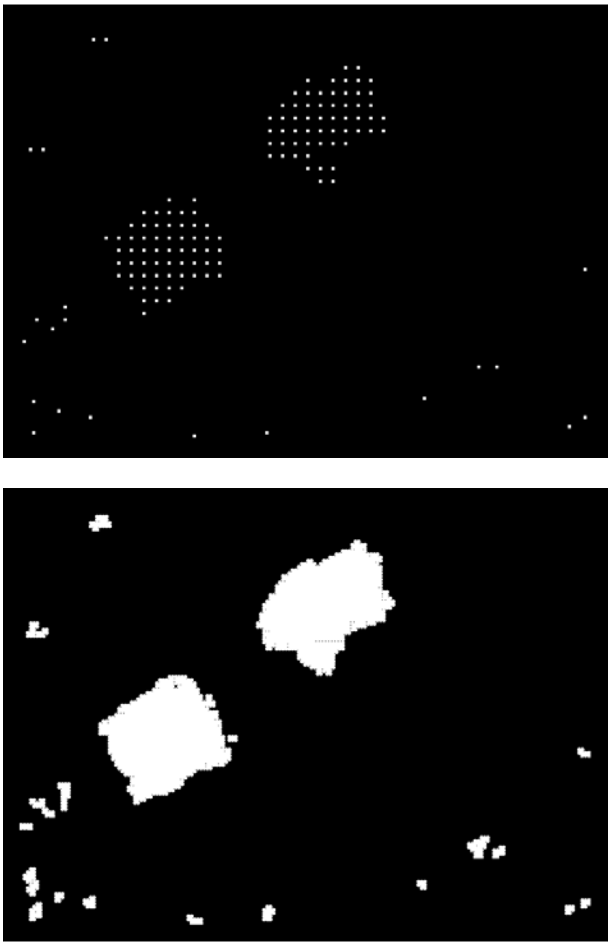
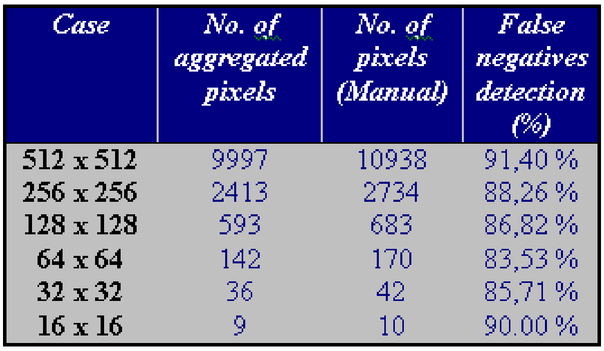
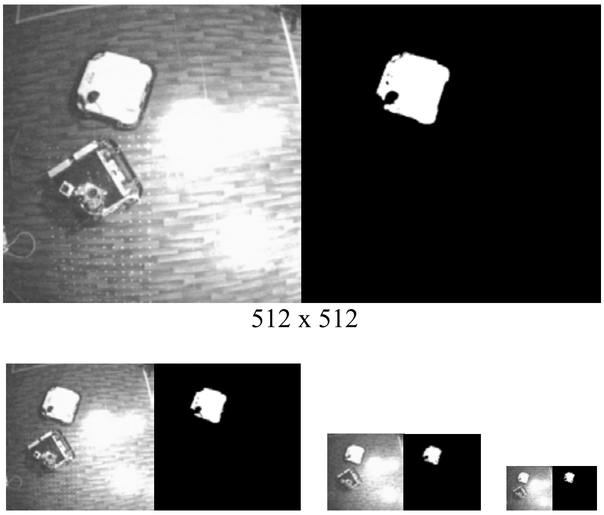


Fig. 5. The sub-sampled foreground pixels obtained with preprocessing filter (top) and the segmented image (bottom).

To get a quantitative evaluation of that interesting property of the SRG method, the false negative detections rate has been calculated for the same image at scales ranging from 512 x 512 pixels to 16 x 16 pixels, representing the same scene at different scales. This is a similar effect as with a discontinuous zoom. It should be noted that the false-negative detection rate is stable regardless of the scale change. The results are demonstrated in the following table:



Case	*No. of aggregated pixels*	*No. of pixels (Manual)*	*False negatives detection (%)*
512 x 512	9997	10938	91,40 %
256 x 256	2413	2734	88,26 %
128 x 128	593	683	86,82 %
64 x 64	142	170	83,53 %
32 x 32	36	42	85,71 %
16 x 16	9	10	90.00 %



512 x 512

256 x 256

128 x 128

64 x 64

Fig. 6. The robustness of the SRG in different scales (512 x 512, 256 x 256, 128 x 128, and 64 x 64).

5 Conclusions

In this paper, we have developed an algorithm for illumination invariant motion detection by combining the adaptive background subtraction algorithm described in [1] with the sparse homomorphic filtering. The adaptive background algorithm deals robustly with slow/normal lighting changes and slow-motion objects. It allows for the introduction or removal of objects from the scene within a sequence (problem of intruders). In our application, results obtained confirm that homomorphic motion detection algorithm is absolutely insensitive to fast changes, up to very fast variations in illumination. It was reported that quick changes in cloud cover disturbed the system described in [1] for 10 to 20 seconds, and the tracking faculty was missing during this period of time. However, introducing the preprocessing (homomorphic filter) eliminated such a drawback completely.

In addition to the illumination invariance, there is another gain achieved by sub-sampling of the image over a rectangular grid. The computational cost and the memory required for the parameters of the five Gaussian distributions, involved to process every pixel in the image as in [1], have been considerably reduced without destroying the efficiency of this approach. As a result of this modification, Stauffer-Grimson method can run in real-time on any typical computer.

The limitation of the presented method is the following: if the moving object is smaller or if the object moves from frame-to-frame less than the distance between the subsamples, then this object will not be detected at all. However, to avoid such situations (supposing that a priori information about the object size is available), the grid size can be made fine enough to "catch" any one of the objects (mobile robots and human beings) in a sampling process. Also, it should be noted that in the presented examples the frame size is 768 x 576 pixels, which is a considerably large one, so a 4 x 4 grid size can be chosen without the risk of missing big blobs (robots and human intruders). The appearance of these blobs depends upon the distance between the camera and the objects on the scene. So the selection of the grid granularity has to account for that as well. A secondary selection factor is related to the fact that the spatial-frequency of sampling strongly affects the processing time. So there is a trade-off between the accuracy of detection and the processing rate.

References

[1] C. Stauffer and W.E.L. Grimson, “Learning patterns of activity using real-time tracking,” in *IEEE Trans. PAMI*, vol. 22, pp. 747-757, Aug. 2000.

[2] D. Toth, T. Aach, and V. Metzler, “Illumination-invariant change detection,” in *Proc. 4th IEEE SouthWest Symp. on Image Analysis & Interpretation*, pp. 3-7, 2000.

[3] R. Baczyk and P. Skrzypczynski, “Mobile robot localization by means of an overhead camera,” in *Proc. Automation 2001*, Warsaw, pp. 220-229, March 2001.

[4] A. Kasinski and A. Hamdy, “Efficient Separation of mobile objects on the scene from the sequence taken with an overhead camera,” in *Proc. Int. Conf. on Computer Vision and Graphics*, Zakopane, vol. 1, pp. 425-430, September 2002.

[5] A. Kasinski and A. Hamdy, “Efficient object segmentation techniques for tracking mobile objects over a sequence of noisy images,” in *Computer Recognition Systems KOSYR 2001*, Wroclaw Univ. of Techn. Press, pp. 421-426, May 2001.

[6] A. Kasinski and A. Hamdy, “Segmentation based on homomorphic filtering and improved seeded region growing for mobile robots tracking in image sequences,” *Machine Graphics & Vision*, vol. 10, no. 4, pp. 447-466, 2001.

[7] D. Anoraganingrum, S. Kroner, and B. Gottfried, “Cell segmentation with adaptive region growing,” in *Proc. 10th ICIAP*, Venice, 1999.

[8] J. S. Lim, *Two-Dimensional Signal and Image Processing*, Prentice Hall, Englewood Cliffs, NJ, pp. 463-465, 1990.

[9] G. Halevy and D. Weinshall, “Motion of disturbances: Detection and tracking of multi-body non-rigid motion,” *Machine Vision and Applications*, pp. 122-137, 1999.

[10] W.E.L. Grimson, C. Stauffer, R. Romano, and L. Lee, “Using adaptive tracking to classify and monitor activities in a site,” in *Computer Vision and Pattern Recognition (CVPR)*, June 1998.

[11] C. Stauffer and W.E.L. Grimson, “Adaptive background mixture models for real-time tracking,” in *Proc. Computer Vision and Pattern Recognition (CVPR)*, June 1999.

[12] D.A.Forsyth and J.Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, pp. 172-175, 2003.

[13] G. Wolberg and H. Massalin, “Fast convolution with packed lookup tables,” *Graphics Gems IV*, Ed. by P. Heckbert, Academic Press, 1994.