Integrated Region- and Pixel-based Approach to Background Modelling

M. Cristani, M. Bicego and V. Murino
Dipart. di Informatica, Università di Verona
Strada Le Grazie 15, 37134 Verona, Italy.
marco.cristani@students.univr.it, bicego@sci.univr.it, murino@sci.univr.it

Abstract

In this paper a new probabilistic method for background modelling is proposed, aimed at the application in video surveillance tasks using a monitoring static camera. Recently, methods employing Time-Adaptive, Per Pixel, Mixture of Gaussians (TAPPMOG) modelling have become popular due to their intrinsic appealing properties. Nevertheless, they are not able per se to monitor global changes in the scene, because they model the background as a set of independent pixel processes. In this paper, we propose to integrate this kind of pixel-based information with higher level region-based information, that permits to manage also sudden changes of the background. These pixel- and region-based modules are naturally and effectively embedded in a probabilistic Bayesian framework called particle filtering, that allows a multi-object tracking. Experimental comparison with a classic pixel-based approach reveals that the proposed method is really effective in recovering from situations of sudden global illumination changes of the background, as well as limited non-uniform changes of the scene illumination.

1 Introduction

Analysis and understanding of video sequences is an active research field which has rapidly increased the interest of the scientific community in the last years, due to the availability of more and more powerful hardware, the development of effective techniques, and the potential vastity of involved applications. Video surveillance is undoubtedly one of the most interesting application of sequence analysis, due to its immediate applicability in several contexts, for instance, the monitoring of parking and working areas, supermarkets, indoor and confined outdoor environments in general.

A videosurveillance system contemplates typically the monitoring of a site for long periods, using a static camera: the goal is to detect and classify moving objects (*foreground*) from static information (*background*). A fundamental issue to be solved is therefore the modelling of the background. Recently, methods employing Time-Adaptive, Per-Pixel, Mixture Of Gaussian (TAPPMOG) have become a popular choice for modelling the background [8, 4]. With these methods, the time evolution of each pixel is considered as a spatial independent process, modelled using a mixture of Gaussians. Each mixture is updated as new observations arrive, while decaying the importance of older observations. At each time step and for each pixel, a subset of Gaussians are considered as background, and current observations that do not match this distribution are labelled as foreground. The attractive properties of TAPPMOGs methods are various: first, they are able to slowly adapt their background model to persistent scene appearance modifications, like the motion of a background object; second, they are quite effective in modelling the relatively simple, but largely repetitive scene appearance changes associated with dynamic objects, like moving foliage; third, they are suitable for real-time implementation.

Nonetheless, these techniques present also some drawbacks. For example, the assignment of a pixel to the background or to the foreground is based on a threshold on the Gaussian mixture, that have to be fixed a priori. Another problem is that they consider each pixel as an independent process without any use of spatial information or, more generically, higher-level information. This problem has been recently addressed by [4], where positive and negative feedbacks from higher level modules have been used to guide low level pixel processes. Moreover, the choice of the learning rate, that determines the "speed" of the self adaption of TAPPMOGS methods to variations of the background, is critical. A high learning rate allows them to adapt rapidly to illumination changes, but does not permit the detection of slowly moving objects, or accentuates the foreground aperture phenomena (i.e., when an uniformly colored object moves, internal pixels could not be detected as foreground [10]). On the other side, a low learning rate permits only slow adaptation, hence in case of a sudden change of the background they find numerous false fore-

ground points for several frames until adaptation is completed.

To face this problem, a non-parametric model for background detection has been proposed in [3]. This work does not model each pixel with a mixture of Gaussians (whose number is fixed), but uses a non-parametric prediction algorithm to estimate the probability density function of each pixel, which is continuously updated to promptly capture the fast variations of the pixel intensity. This technique succeeds to better model the behavior of each pixel, not necessarily constrained to fit a set of fixed Gaussians, but it still requires the use of thresholds to be tuned to get the desired performances (number of false positives).

The sensitivity to global changes of the illumination of the scene is another delicate issue. This is one of the most severe problems to be solved by a background modelling system, especially if changes are local and not uniformly distributed over the scene. Actually, a variation on the illumination of the whole scene could be detected and recovered with a standard histogram normalization technique, whereas local variations could not be detected with such a global analysis. This situation can be very frequent in indoor situations, for example when the door of a lit room is opened in a monitored dark corridor. A substantial contribution in this sense was provided by [9], where a topology free Hidden Markov Model was used in order to model illumination changes of the scene. Even if results are promising, this method does not work on-line, and illumination changes have to be pre-classified off-line. Another interesting approach was proposed by Ohta in [7], where the possible changes in illumination are coded explicitly in a mathematical model. Nevertheless, the effectiveness of the method depends on the number of background prototypes estimated, failing for unexpected illumination changes.

In this paper a novel approach is proposed, which is able to deal with sudden variations of illumination in the scene, also restricted to partial parts of it. We start from a generic TAPPMOG method like that proposed by Stauffer and Grimson [8]. The basic idea of our approach is that this process can be improved if we consider also a sort of region-based modelling, i.e., considering the spatial information as provided by a classical image segmentation. With high probability, a change in illumination, even if restricted to a particular area, results in a variation of the gray-level values of most of the pixels of the regions in that zone. In other words, if all pixels of a region significantly vary simultaneously, a typical system will tend to identify them as foreground, but, if the region is enough large, there is a high probability that this situation can be due to an illumination change rather than actual foreground.

Our approach uses spatial information resulting from a (off-line) spatial segmentation of the background (obtained for example by processing the first frame) as prior in order to modulate the response of a TAPPMOG system. In particular, a variation of the learning parameter of the system is devised in order to cope efficiently with sudden changes in the background appearance.

Subsequently, this approach is naturally integrated in a probabilistic Bayesian framework, the particle filtering [5, 2] paradigm for tracking. This Monte Carlo technique [2], that has recently received growing attention, is based on sequential importance sampling/resampling, and provides a sound statistical framework for propagating sample-based approximations of posterior distributions, with almost no restriction on the ingredients of the model. We will show how a TAPPMOG module can be naturally inserted in this framework, eliminating the mixture threshold problem discussed above. We will also show, on a real sequence available in the literature [9], that the use of spatial information is able to correct and manage sudden changes of illumination, even if restricted to local scene areas.

The rest of the paper is organized as follows. In Section 2, the basics of TAPPMOG-based methods and particle filtering is introduced, mainly to fix the notation. Section 3 details the proposed approach which includes both region and pixel information in a sound statistical model, and, in Section 4, some results are reported, showing a comparison with the classical method. In Section 5, conclusions are finally drawn and future perspectives are envisaged.

2 Fundamentals

2.1 The TAPPMOG background modelling

In this subsection, standard time adaptive per-pixel mixture of Gaussians background modelling scheme is presented, following [8]. A mixture of Gaussians is associated to each pixel, modelling the evolution of its gray level during time. The probability to observe the value $z_{uv}^{(t)}$, i.e., the intensity gray level of the pixel (u, v) of the image at time t, is given by:

$$P(z_{uv}^{(t)}) = \sum_{j=1}^{K} w_{j,uv}^{(t)} \mathcal{N}\left(z_{uv}^{(t)} | \mu_{j,uv}^{(t)}, \sigma_{j,uv}^{(t)}\right) \qquad (1)$$

where $w_{j,uv}^{(t)}$, $\mu_{j,uv}^{(t)}$ and $\sigma_{j,uv}^{(t)}$ are respectively the mixing coefficients, the mean and the standard deviation of the j-th Gaussian of the mixture of the pixel (u, v) at time t. The background modelling algorithm proceeds as follows. Suppose that, at each time instant, the Gaussians in a mixture are ranked in descending order by the value of w/σ. Every new pixel value is checked against the existing K Gaussian functions until a match is found, where a success match is defined as a pixel value within 2.5σ of any mode of the distribution. If none of the K Gaussian functions matches the

pixel value, the least probable function is replaced with a new one, having mean equal to the current value, high variance, and low mixing coefficient. If j_{hit} is the Gaussian component matched, a pixel $z_{uv}^{(t)}$ is labelled as foreground if

$$\sum_{j=1}^{j_{hit}} w_{j,uv}^{(t)} > T \tag{2}$$

where T is a threshold (to be defined a priori) that indicates the minimum portion of the data that should be accounted for by the background.

Each mixture evolves during time, as new evidence arrives. For the mixing coefficients:

$$w_{j,uv}^{(t)} = (1-\alpha) w_{j,uv}^{(t-1)} + \alpha M_{uv}^{(t)}, 0 \leq j \leq K, \tag{3}$$

where $M_{uv}^{(t)}$ is 1 for the matched Gaussian and 0 for the others, and α is the learning rate. Low α values imply a slow adaption, and vice versa. The μ and σ parameters for unmatched Gaussians remain the same, but, for the matched Gaussian function j_{hit} we have (omitting indexes for clarity):

$$\mu^{(t)} = (1-\rho)\mu^{(t-1)} + \rho z^{(t)} \tag{4}$$

$$\sigma^{2\,(t)} = (1-\rho)\sigma^{2\,(t-1)} + \rho \left(z^{(t)} - \mu^{(t)}\right)^T \left(z^{(t)} - \mu^{(t)}\right) \tag{5}$$

where $\rho = \alpha \mathcal{N}\left(z_{uv}^{(t)} | \mu_{j,uv}^{(t)}, \sigma_{j,uv}^{(t)}\right)$.

2.2 The particle filtering tracker

Due to lack of space, a comprehensive description of this approach is not presented here, and only the general ideas are introduced, mainly to setup the notation. Interested readers are referred to [5, 2, 6].

The particle filtering is a Bayesian approach, assuming that all information obtainable from the image Z^t about the model X^t, which represents the moving (foreground) objects in the scene at time t, is encoded in the posterior distribution $P(X^t|Z^t)$. This probability is approximated using a set of samples $\{s_{(\ell)}^t, \pi_{(\ell)}^t\}$, where each sample represents an instance of the model X^t with a probability $\pi_{(\ell)}^t$. The algorithm, in its general formulation, follows a set of rules for propagating this set of samples over time. Basically, at each time instant t, the following steps are performed:

- *sampling from prior (the posterior of step $t-1$)*: L samples are chosen from $\{s_{(\ell)}^{t-1}\}$ with probability $\{\pi_{(\ell)}^{t-1}\}$, obtaining $\{\tilde{s}_{(\ell)}^{t-1}\}$. In this way, samples with high weight at time t-1 have higher probability to "survive".
- *prediction*: samples $\{s_{(\ell)}^t\}$ for time t are then obtained by applying a dynamics to $\{\tilde{s}_{(\ell)}^{t-1}\}$, predicting the new configurations based on previous values and on some a priori knowledge about the possible movements of the objects; typically, this dynamics also contains a stochastic component.
- *weighting*: samples obtained by previous step are then weighted using the evidence $P(Z^t|X^t)$ (also called *likelihood*) from the image Z^t; at each sample $s_{(\ell)}^t$ is then assigned the weight $\pi_{(\ell)}^t$, computed as $\pi_{(\ell)}^t = P(Z^t|X^t = s_{(\ell)}^t)$.

At each time step t, the estimated model X^t could be obtained with a MAP approach, i.e. by choosing the most probable sample.

In our approach, we did not use the pixel as elementary image entity, but the response of a circular Gaussian filter of mean 0 and variance 1. These filters are spaced in the image every 5 pixels, and are partially overlapped. The set of the responses of each filter at time t yields the "image" $Z^t = \{z_n^{(t)}\}$.

The definition of the sample $s_{(\ell)}^t$ follows the idea of multiple blob tracker proposed in [6]: $s_{(\ell)}^t$ is a configuration $\left(m_\ell^t, x_{\ell,1}^t, x_{\ell,2}^t, ..., x_{\ell,m_\ell^t}^t\right)$, where m_ℓ^t is the number of objects, and $x_{\ell,i}^t$ are the positions of the objects in the scene. Each object is simply described with a vertically oriented ellipse, centered on $x_{\ell,p}^t$, called $\mathcal{E}(x_{\ell,p}^t)$. The dynamics (second step of the algorithm) operates on the samples by processing not only the objects' positions, but also the number of objects. In this way, the system is also able to track several objects, managing also entities which are entering or exiting in the scene. Finally, the likelihood of a configuration $s_{(\ell)}^t$ is computed starting from the background response $L(z_n^{(t)}) = P(z_n^{(t)} \in FG)$, that represents the probability that the filter n is foreground at time t. The likelihood is zero for the configurations in which not all ellipses are covered by a sufficient foreground. For the others, the likelihood is computed as:

$$P(Z^t|X^t = s_{(\ell)}^t) = \frac{1}{k}\left(\left(\sum_p \sum_{n \in \mathcal{E}(x_{\ell,p}^t)} L(z_n^{(t)})\right) - \sum_{n \notin \mathcal{E}(x_{\ell,p}^t)} L(z_n^{(t)})\right) \tag{6}$$

where k is a normalization constant. In other words, a positive contribution to the likelihood of the sample $s_{(\ell)}^t$ derives from filters "covered" by the objects of $s_{(\ell)}^t$, whereas the others filters contribute negatively. In this way, the configurations that correctly predict both the positions and the number of objects in the scene have higher likelihood than configurations that correctly predict only the positions of a less number of objects.

3 The integrated region and pixel-based approach

In this section, the proposed approach is detailed: first, we describe how a TAPPMOG-based system is extended to naturally incorporate spatial information and to be encapsulated in the particle filtering framework; second, we explain the strategy that uses region-based information to modulate the pixel-based response, in order to obtain the background response L needed by the tracking algorithm.

The starting point is a spatial segmentation of the background scene (for instance, obtained by segmenting, using a region growing approach [1], the first frame of the sequence). The segmented image is defined as $R = \{R_i\}$, $1 \leq i \leq M$, and $R_i = \{R_i^1 ... R_i^{|R_i|}\}$, where $|R_i|$ is the size of region R_i and R_i^n is the n-th filter of the region R_i. We denote as $z_n^{(i,t)}$ the observation of the n-th filter of the i-th region at time t.

The unmodulated pixel-level background response $\tilde{L}(z_n^{(i,t)})$ is naturally obtained by computing

$$\tilde{L}(z_n^{(i,t)}) = P\left(z_n^{(i,t)} \in \text{ FG}\right) = \sum_{j=1}^{j_{hit}} w_{j,n}^{(i,t)} \qquad (7)$$

representing the probability that $z_n^{(i,t)}$ is foreground, which is assigned by the TAPPMOG module, i.e. before high-level modulation. The weights $w_{j,n}^{(i,t)}$ are mixing coefficients related to the j-th Gaussian of the mixture corresponding to the n-th filter of the i-th region, at time t. It is worth noticing that in this way the threshold T, present in the Eq. (2), is not required anymore. The tracking algorithm uses all information embedded in Eq. (7), without any loss derived from the thresholding approximation.

Subsequently, the spatial information derived from segmentation is used to modulate the low-level response, varying the learning parameter α in order to allow the system to rapidly evolve in case of sudden change of the background. The idea is to "accelerate", when needed, the process of adaptiveness of the low level module. For example, with a sudden change in illumination, the most of pixels of the interested region changes suddenly, thus obtaining a wrong high probability to be foreground. Monitoring these sudden changes, we can adapt learning parameters in order to recover from these situations. To do that, we define for each region R_i the *approximate filling coefficient* $\gamma_i^{(t)}$, that represents the probability, assigned by the low level module, that a region R_i is foreground:

$$\gamma_i^{(t)} = \frac{\sum_{n=1}^{|R_i|} \tilde{L}(z_n^{(i,t)})}{|R_i|} \qquad (8)$$

We define also the *modulated filling coefficient* $\hat{\gamma}_i^{(t)}$ in the same manner, only using $L(z_n^{(i,t)})$ instead of $\tilde{L}(z_n^{(i,t)})$. $L(z_n^{(i,t)})$ represents the final estimate, after modulation, of the probability of being foreground of the n-th filter of the region R_i. The computation of this quantity is described later in this section.

Instead of having a fixed learning parameter α, we propose to have, at each time step t, a set of learning parameters $\alpha_i^{(t)}$, one for each region R_i. These coefficients are computed with the following formula:

$$\alpha_i^{(t)} = \max\left(\alpha, \left|\gamma_i^{(t)} - \hat{\gamma}_i^{(t-1)}\right|\right) \qquad (9)$$

where α is the TAPPMOG learning parameter of formulas (3) and (4): this was fixed to 0.7, value that permits to detect also relatively slowly moving objects.

The quantity $\left|\gamma_i^{(t)} - \hat{\gamma}_i^{(t-1)}\right|$ represents a measure of how much part of the region R_i is changed from step $t-1$ to step t. If this quantity is low, the low-level module does not need rectification or adjustments. On the contrary, when this quantity is high, a large part of the region R_i has changed rapidly, and, if the regions are sufficiently larger than the foreground, this rapid change can be likely due to an illumination variation. In this case, the background model must adapt very fast to this new situation, hence the learning parameter should be increased. Moreover, this upsurge of the speed of adaptiveness is not a priori fixed, but depends on the rapidity and the globality of the background change.

The increase of the adaptiveness speed means that, in the update of the parameters, most of the importance is given to the last observation (the one of the illumination change), forcing it to become rapidly one of the background Gaussians. This is correct if the whole region is background, but, if foreground is present during the change, this update is indeed wrong. In this case, the algorithm sets as background what is actually foreground, losing the foreground in the scene. This is solved by using, in the updating parameter equations (Eq. (3) and (4)), the value $\hat{z}^{(i,t)}$ instead of $z_n^{(i,t)}$. This value is the weighted average of the observations $z_n^{(i,t)}$ of the filters of the region R_i, each weighted by its probability to be background at time step $t-1$, i.e.,

$$\hat{z}^{(i,t)} = \frac{1}{k} \sum_{n=1}^{|R_i|} (1 - L(z_n^{(i,t-1)})) z_n^{(i,t)} \qquad (10)$$

where k is a normalization constant. In this way, the system is able to detect the foreground also after the reparameterization of the background model. The use of this averaged region-based value to update the model, instead of using pixel-based (or filter-based) value, is actually reasonable in that the segmentation used as prior knowledge determines regions of gray-level similarity. Consequently, by substituting each value in the region with the averaged region value results in a approximation, indeed sufficient to

recover from illumination change situations. This region-based approximation is then refined in a couple of frames by the usual time-adaptation of the TAPPMOG pixel-based process.

If the learning parameter $\alpha_i^{(t)}$ is changed, the mixture parameters of the whole region are adjusted accordingly. From the update, we obtain new mixture parameters' estimates, $\hat{w}_{j,n}^{(i,t)}, \hat{\mu}_{j,n}^{(i,t)}, \hat{\sigma}_{j,n}^{(i,t)}$, and we re-compute the likelihood L, allowing an immediate correction and recovery from illumination changes. The final modulated background response L, used by the tracker (Eq. (6)), is finally computed as:

$$L(z_n^{(i,t)}) = \begin{cases} \tilde{L}(z_n^{(i,t)}) & \text{if } \alpha_i^{(t)} = \alpha \\ \sum_{j=1}^{\hat{j}_{hit}} \hat{w}_{j,n}^{(i,t)} & \text{otherwise} \end{cases} \quad (11)$$

4 Results

The proposed approach was compared with [8] using different sequences, presenting or not illumination changes. In the former case, no relevant differences were found between the two approaches; more interesting is the latter case, where the non uniform illumination change drastically affects the TAPPMOG performances. One of such sequences was obtained from [9], regarding the monitoring of an indoor environment with one moving object . The sequence is formed by 160 frames, acquired at 20 frame/sec. Some of the frames of the sequence are presented in Fig. 1, showing the illumination change, occurring at frames 83-84. The initial spatial segmentation used in this experiment is shown in Fig. 2. In Fig. 3, a comparison between standard TAPPMOG method as in [8] and the proposed approach is presented (white pixels represent the foreground). We can notice that, in correspondence of the sudden change of illumination (frames 83-84), the TAPPMOG method identifies almost all pixels in the scene as foreground. This is obvious as the per pixel process recognizes only the pixel gray level variation. With our approach, the use of the spatial high-level information permits the detection of the globality of the change, recovering in real time the correct background. We can also notice that when the foreground object actually comes in again in the scene at frame 100, our approach succeeds to distinguish it (right of the images of the right column of Fig. 3), whereas the TAPPMOG method succeed to discriminate it after a certain latency, only at frame 112. More precisely, the TAPPMOG approach needs 28 frames for the adaptation to the change of illumination, whereas in the proposed approach the recover is immediate. This is confirmed by results obtained applying tracking procedure, proposed in figure 4. We could notice that, before the illumination change, the object (identified by the red ellipse) is correctly tracked by both methods. After the change, instead, the background response given by the TAPPMOG

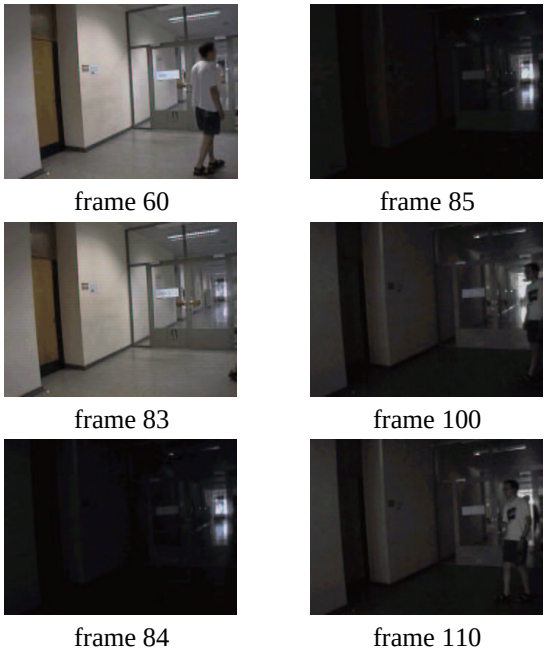

Figure 1. Frames from the test sequence.

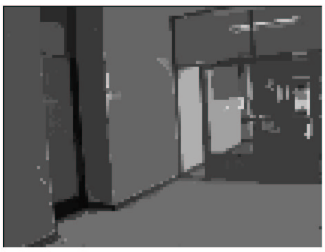

Figure 2. Spatial segmentation of the background.

module is not-informative, and the tracker is not able to detect the incoming object. With our approach, instead, the response of the background module is correct, and the object is correctly tracked. It is worthwhile noticing that the obtained results are similar to those proposed in [9]: nevertheless, our approach, as [8], could be executed in real time, whereas [9] runs off-line.

5 Conclusions

In this paper, a novel method for background modelling is proposed. The idea is to modulate pixel-based information with higher level region-based information, represented

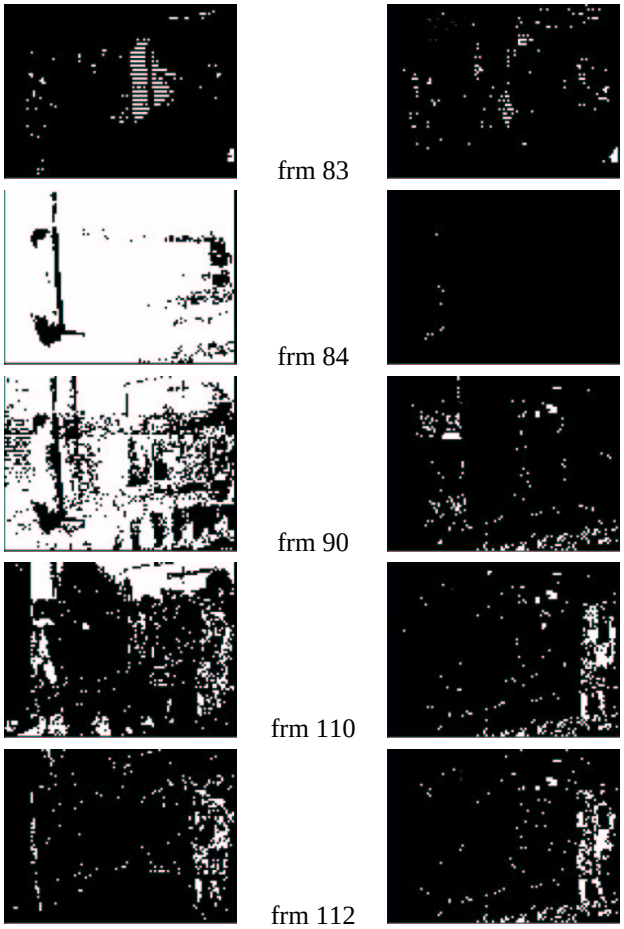

Figure 3. Response of the background model: (left) standard TAPPMOG module; (right) the proposed approach.

by a spatial segmentation of the background scene. This modulation results in a variation of the adaptiveness speed of the background modelling system driven by region-based reasoning. The presented system is naturally and effectively integrated in a probabilistic multi-object tracking framework, namely, the particle filtering, which allows a seamless management of the available information, and avoids the use of heuristic thresholds (indeed utilized in the classic approach). Experimental results have shown that our approach is able to effectively recover from sudden changes in the illumination of the scene.

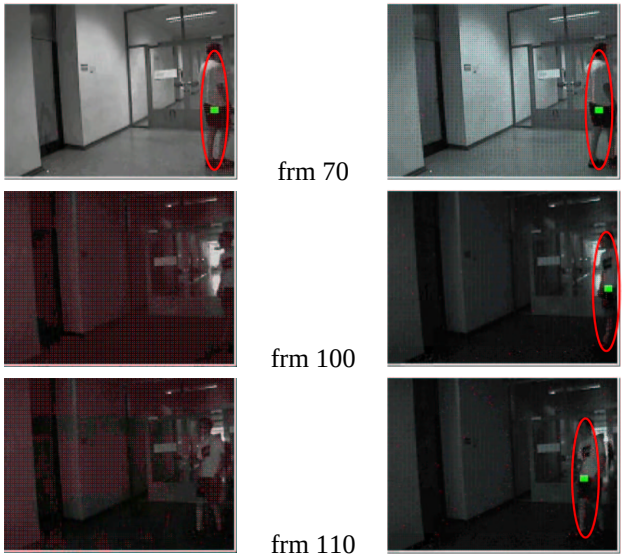

Figure 4. Tracking of the foreground, using different background modules: (left) standard TAPPMOG; (right) the proposed approach.

References

[1] K. Castleman. *Digital Imgage Processing*. Prentice Hall, 1996.

[2] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[3] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background substraction. In *European Conf. Computer Vision*, 2000.

[4] M. Harville. A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models. In *European Conf. Computer Vision*, volume 3, pages 543–560, 2002.

[5] M. Isard and A. Blake. CONDENSATION: Conditional density propagation for visual tracking. *Int. J. of Computer Vision*, 29(1):5–28, 1998.

[6] M. Isard and J. MacCormick. BraMBLe: a bayesian multiple-blob tracker. In *Int. Conf Computer Vision*, volume 2, pages 34–41, 2001.

[7] N. Ohta. A statistical approach to background subtraction for surveillance systems. In *Int. Conf. Computer Vision*, volume 2, pages 481–486, 2001.

[8] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Int. Conf. Computer Vision and Pattern Recognition*, volume 2, 1999.

[9] B. Stenger, V. R. nad N. Paragios, F.Coetzee, and J. M. Buhmann. Topology free hidden markov models: Application to background modeling. In *Int. Conf. Computer Vision*, volume 1, pages 294–301, 2001.

[10] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Int. Conf. Computer Vision*, pages 255–261, 1999.