Adaptive Model for Object Detection in Noisy and Fast-Varying Environment

Dung Nghi Truong Cong[1], Louahdi Khoudour[1],
Catherine Achard[2], and Amaury Flancquart[1]

[1] IFSTTAR, LEOST, F-59650 Villeneuve d'Ascq, France
[2] UMPC Univ Paris 06, ISIR, UMR 7222, France
truong@ifsttar.fr, louahdi.khoudour@ifsttar.fr,
catherine.achard@upmc.fr, amaury.flancquart@ifsttar.fr

Abstract. This paper presents a specific algorithm for foreground object extraction in complex scenes where the background varies unpredictably over time. The background and foreground models are first constructed by using an adaptive mixture of Gaussians in a joint spatio-color feature space. A dynamic decision framework, which is able to take advantages of the spatial coherency of object, is then introduced for classifying background/foreground pixels. The proposed method was tested on a dataset coming from a real surveillance system including different sensors installed on board a moving train. The experimental results show that the proposed algorithm is robust in the real complex scenarios.

Keywords: Background subtraction, foreground segmentation, mixture of Gaussians, spatio-color feature space.

1 Introduction

Detecting foreground objects from a video sequence is a critical task in many computer-vision applications. It can be considered as the basic level of processing to achieve higher level vision tasks. Even though there exist numerous algorithms in the literature, foreground object detection in complex environments, including non-stationary background motion, illumination variations, and camera vibration, is still far from being completely solved.

As surveyed in [1], there exists a vast literature on background subtraction. Most proposed methods are based on the pixel-level background model, which construct a background representation for each pixel location. One of the simplest approaches consists in modeling each pixel intensity with a single Gaussian distribution [2]. However, such a model is unsuitable for noisy sequences and multi-modal scenes. More complex models are based on a mixture of Gaussians [3], or a probability density function estimated by kernel function [4]. The background can also be modeled by a group of clusters which represent a compressed form of background model [5].

In contrast to pixel-wise approach, interest has grown recently in region-level methods which employ regional models representing spatial relationships between pixels. Sheikh et al. [6] used Kernel Density Estimation to build full background model as a single distribution, in conjunction with a MAP-MRF decision framework. In [7], Heikkila et al. used a group of weighed adaptive local binary pattern histograms to capture the

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part I, LNCS 6978, pp. 68–77, 2011.

background statistics of each image block, and produced a coarse detection of foreground object. Chen et al. [8] extended this idea to obtain more detailed foreground by using a contrast histogram to describe each block. More recently, Dickinson et al. [9] modeled the background as an adaptive mixture of Gaussians in color and space, and used this model to probabilistically classify new pixels observations.

In this article, we consider the problem of foreground object detection in a complex environment where the background varies unpredictably over time. This work is carried out in the framework of the BOSS European project (on BOard wireless Secured video Surveillance), whose objective is to set up an onboard surveillance system. Indeed, the complex environments inside a moving train make the detection task extremely difficult. Therefore, in order to deal with such particular problems, we propose an approach based on an adaptive spatio-colorimetric background and foreground model coupled with a dynamic decision framework. The proposed method has three novel contributions. Firstly, in order to handle multi-modal uncertainties of the background, a joint spatio-colorimetric region based representation is employed to model the observed scene. The statistical regions of the background, which share common homogeneity properties, are modeled by using an adaptive mixture of Gaussians in a five-dimensional spatio-colorimetric feature space. Secondly, both background and foreground are modelized in order to better distinguish foreground and background pixels. Thirdly, instead of directly applying a threshold to classify background/foreground pixels, we propose a dynamic decision framework based on cellular automata which enforces the spatio-colorimetric context in the detection process.

The outline of the paper is as follows: after this introduction, we present in Section 2 the proposed approach to extract foreground objects. Section 3 presents global performances of the proposed system on different real datasets. Finally, in Section 4, conclusions and important short-term perspectives are given.

2 The Proposed Approach

2.1 Modeling the Background

The initial representation of the background is constructed from the first frame of the sequence by using a region merging technique [10] coupled with an adaptive mixture of Gaussians. The homogeneous regions of the observed scene are first extracted by iteratively combining smaller pixels or regions sharing homogeneous color properties.

Let I be the observed image containing N pixels; (p, p') be a couple of adjacent pixels in 4-connexity and A_I be the set of these couples. We first compute the local gradient between each couple of pixels defined as:

$$g\left(p, p'\right) = \max_{c \in \{R,G,B\}} \left| \overline{R_p(p')_c} - \overline{R_{p'}(p)_c} \right| \tag{1}$$

where $R_p\left(p'\right)$ is the set of neighborhood pixels of p' which satisfies the condition: $R_p\left(p'\right) = \left\{q \in I : \|q - p'\|_1 < \delta \;\&\; \|q - p'\|_1 < \|q - p\|_1\right\}$ (δ is a predefined radius depending on the noise corruption of images; it is set to $\delta = 5$ in our experimentations), $\overline{R_p(p')_c}$ is the mean color value of channel c of all pixels belonging to $R_p\left(p'\right)$. The principle of region merging process is that the couples of A_I are first sorted in increasing order of $g\left(p, p'\right)$. For each couple of pixels $\left(p, p'\right) \in A_I$, let $r\left(p\right)$ and $r\left(p'\right)$ be the current regions to which pixels p and p' belong respectively. These two regions are merged if the following condition is verified:

$$\left|\overline{r(p')_c} - \overline{r(p)_c}\right| \leq \kappa\sqrt{\frac{1}{N_{r(p)}} + \frac{1}{N_{r(p')}}}\ , \ \forall c \in \{R, G, B\} \tag{2}$$

where $\overline{r(p)_c}$ is the mean color value of the channel c of region $r\left(p\right)$, $N_{r(p)}$ is the number of pixels of region $r\left(p\right)$ and κ is a parameter defined as:

$$\kappa = C\sqrt{\frac{2\log\left(N\right)}{\Phi}} \tag{3}$$

where C is the maximum value of color space, N is the number of pixels of the observed image and Φ is a parameter modifying the coarseness of the segmentation.

The region merging process is resumed in Algorithm 1.

Algorithm 1. Region merging algorithm

```
Initialization: the set $A_I$ and the local gradient value of each couple $g\left(p, p'\right)$.
Sorting the couples of $A_I$ in increasing order of $g\left(p, p'\right)$
for each couple $\left(p, p'\right) \in A_I$, $r\left(p\right) \neq r\left(p'\right)$ do
    if $\left|\overline{r(p')_c} - \overline{r(p)_c}\right| \leq \kappa\sqrt{\frac{1}{N_{r(p)}} + \frac{1}{N_{r(p')}}}\ \forall c \in \{R, G, B\}$ then
        merging $r\left(p\right)$ and $r\left(p'\right)$
```

After this region merging procedure, the observed scene is segmented into K_B homogeneous regions. Each region is now modeled by a Gaussian distribution in the joint spatio-colorimetric feature space $\mathbf{x} = \left[x, y, R, G, B\right]^T$ where x and y are spatial coordinates in the two-dimensional image and R, G, B are color coordinates:

$$\eta\left(\mathbf{x}|\mu, \Sigma\right) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}\left(\mathbf{x} - \mu\right)\right)}{\sqrt{\left(2\pi\right)^5\left|\Sigma\right|}} \tag{4}$$

Here μ and Σ are the mean vector and covariance matrix estimated on the pixels belonging to the current region. The background model is finally defined by:

$$f\left(\mathbf{x}|BG\right) = \sum_{i=1}^{K_B} w_i \eta\left(\mathbf{x}|\mu_i, \Sigma_i\right) \tag{5}$$

where $w_i = N_i/N$ is the weight of the i^{th} component, N_i is the number of pixels of the region i.

2.2 Modeling the Foreground

Since foreground objects tend to have smooth motion from frame to frame, the temporal persistence property is a powerful tool to increase the accuracy of object detection. Here, we propose to model the foreground objects in order to employ simultaneously background and foreground models to improve the detection.

The foreground model is initialized as a uniform function. Once a foreground region is detected, the foreground model is constructed in the same manner as the background one and is expressed by:

$$f\left(\mathbf{x}|FG\right) = \alpha + (1-\alpha)\sum_{i=1}^{K_F} w_i^F \eta\left(\mathbf{x}|\mu_i^F, \Sigma_i^F\right) \tag{6}$$

where α is a constant which yields robustness when foreground is not observed ($\alpha < 0.5$), K_F is the number of components of the foreground model.

2.3 Foreground Object Segmentation

In this section, we propose a particular algorithm for foreground object segmentation based on the principle of cellular automata introduced by von Neumann [11]. The idea of this algorithm is that each pixel p is considered as a cellular automaton characterized by a triplet (l_p, N_p, Δ), where l_p and N_p are respectively the label and the set of neighborhood pixels of the current pixel p, Δ is the local transition function. The pixel label at instant $k+1$ is estimated based on the states of the neighborhood pixels at instant k and the local transition rule.

Each pixel p of the new captured frame is first classified into one of three classes (foreground, background or undefined) based on the likelihood ratio $\Gamma = -\log\dfrac{f\left(\mathbf{x}_p|BG\right)}{f\left(\mathbf{x}_p|FG\right)}$. The label of pixel p is defined as:

$$l_p = \begin{cases} -1 \;\; (BG) & \text{if } \; \Gamma < T_{BG} \\ 1 \;\; (FG) & \text{if } \; \Gamma > T_{FG} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where T_{BG} and T_{FG} are two parameters a priori defined for classifying foreground and background pixels.

The confidence score of each pixel is also estimated by using the probability of observing a background/foreground pixel:

$$\begin{cases} C_p = f\left(\mathbf{x}_p|BG\right) & \text{if } \; l_p = -1 \\ C_p = f\left(\mathbf{x}_p|FG\right) & \text{if } \; l_p = 1 \\ C_p = 0 & \text{if } \; l_p = 0 \end{cases} \tag{8}$$

Algorithm 2 describes the foreground object segmentation procedure.

Algorithm 2. Foreground object segmentation

```
$k = 0$: $l_p^0 = l_p$ and $C_p^0 = C_p$ for all $p \in I$
while not converged do
    for each pixel $p \in I$ do
        $l_p^{k+1} = l_p^k$, $C_p^{k+1} = C_p^k$
        for each pixel $q \in N\left(p\right)$ do
            if $\varDelta\left(p, q\right).C_q^k > C_p^k$ then
                $l_p^{k+1} = l_q^k$
                $C_p^{k+1} = \varDelta\left(p, q\right).C_q^k$
    $k = k + 1$
```

Here, the local transition function $\varDelta$ is defined as:

$$\varDelta\left(p, q\right) = 1 - \exp\left(\frac{-\beta}{\varepsilon + \left\|\mathbf{I}_p - \mathbf{I}_q\right\|_2}\right) \tag{9}$$

where β and ε are the predefined parameters, $\mathbf{I}_p$ and $\mathbf{I}_q$ are the color vectors of pixels p and q.

Thus, instead of applying a single threshold to classify background/foreground pixels, we try to exploit the spatial coherency of object in order to obtain an optimal segmentation. Each pixel is represented by its confidence score. The higher the confidence score C_p, the stronger the influence on the neighborhood pixels. If two neighbor pixels have similar color, $\varDelta\left(p, q\right)$ is big and the label of pixel with lower score will be replaced.

2.4 Updating the Background Model

The background model is updated by first assigning the new background pixels to their corresponding components, and then re-estimating the component parameters. The pixel $\mathbf{x}$ is assigned to component C if:

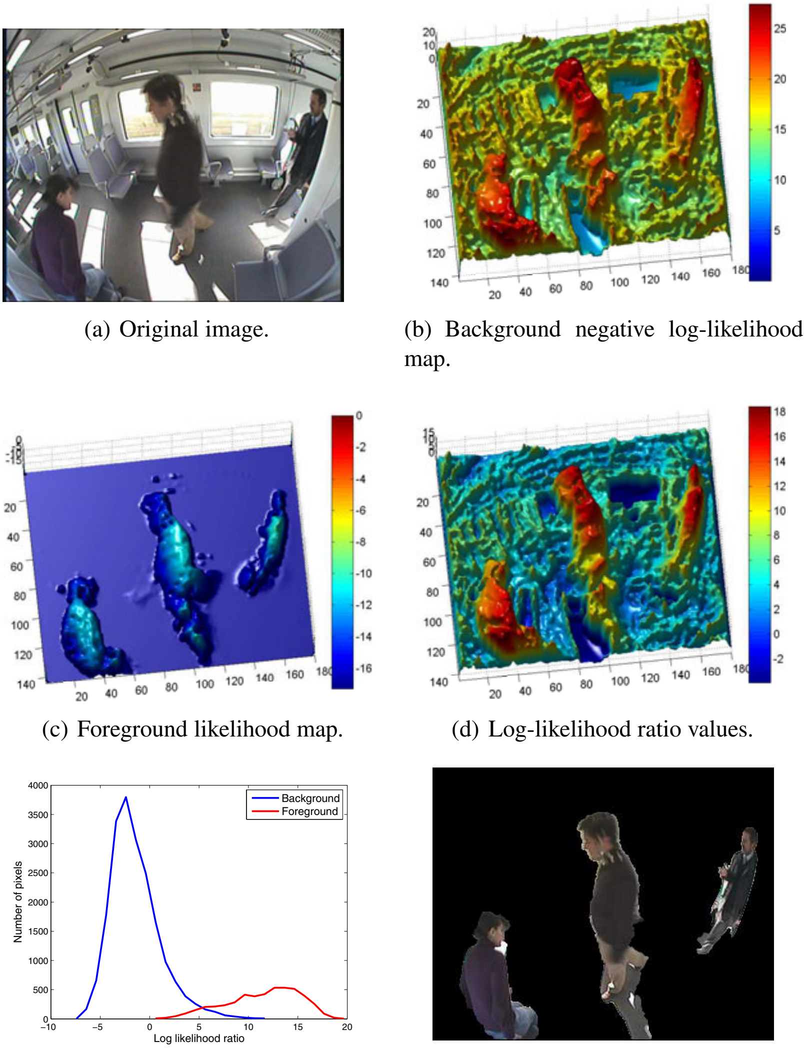
$$C = \arg\max_i \left\{\eta\left(\mathbf{x}|\mu_i, \varSigma_i\right)\right\} \tag{10}$$

Let μ_i^* and $\varSigma_i^*$ be the mean vector and covariance matrix estimated from the N_i^* new pixels assigned to component i. The parameters of component i are re-estimated with:

$$\begin{aligned} w_i^t &= \frac{w_i^{(t-1)}N + N_i^*}{N + N^*} \\ \mu_i^t &= \frac{w_i^{(t-1)}N\mu_i^{(t-1)} + N_i^*\mu_i^*}{Nw_i^{(t-1)} + N_i^*} \\ \varSigma_i^t &= \frac{w_i^{(t-1)}N\varSigma_i^{(t-1)} + N_i^*\varSigma_i^*}{Nw_i^{(t-1)} + N_i^*} - \mu_i^t[\mu_i^t]^T + \frac{w_i^{(t-1)}N\mu_i^{(t-1)}\left[\mu_i^{(t-1)}\right]^T + N_i^*\mu_i^*[\mu_i^*]^T}{Nw_i^{(t-1)} + N_i^*} \end{aligned} \tag{11}$$

Unlike the background, the foreground model is reconstructed by using the new extracted foreground objects in the same manner as initializing the background model. Thus, the foreground model adapts rapidly from frame to frame, which makes the detection task in the next frame more robust.

Figure 1 presents the detection results for an image of the sequence. Images 1(b) and 1(c) represent the likelihood maps for both background and foreground while image 1(f) is the final result of detection.



(a) Original image.

(b) Background negative log-likelihood map.

(c) Foreground likelihood map.

(d) Log-likelihood ratio values.

(e) Histograms of log-likelihood ratio values.

(f) Final result.

Fig. 1. Different steps of the proposed algorithm for foreground object extraction

3 Results and Discussion

The performance of the proposed method is evaluated using the real dataset collected by the cameras installed on board a moving train in the framework of the BOSS European project [12]. This dataset is really difficult, since the captured video is influenced by many factors including fast illumination variations and non-static background due to the movement of the train, reflections, vibrations of the cameras...

Figure 2 illustrates the effectiveness of the proposed method in comparison with the well-known Mixture of Gaussians method (GMM). The first row is the original images, the second row shows the results obtained by GMM method, and the third row presents the results obtained by the proposed method. Note that no post-processing is used in the results.

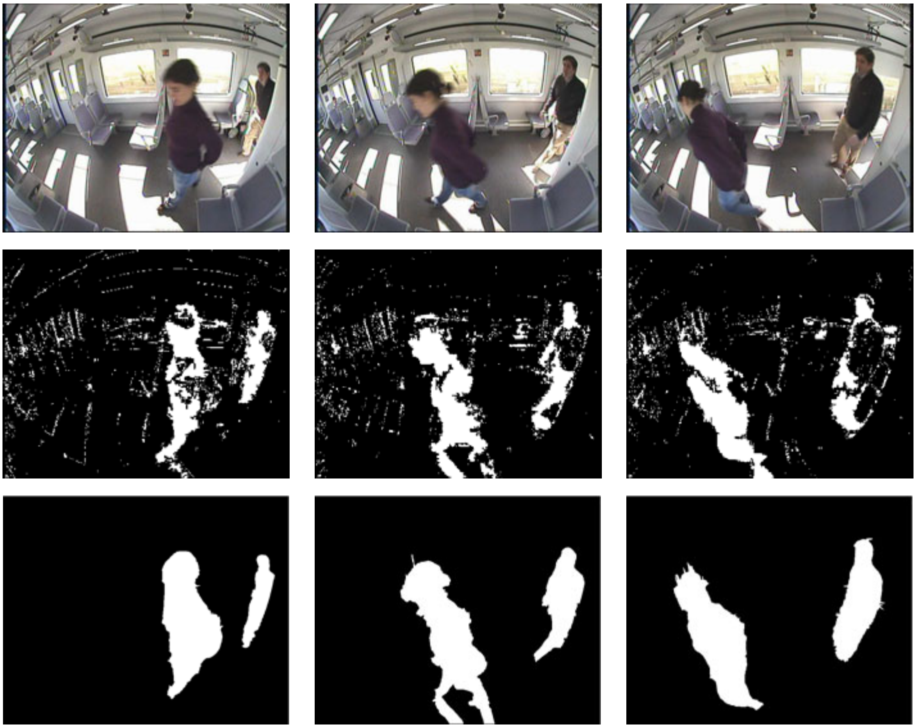


Fig. 2. Foreground object extraction results. Top to bottom: original images, GMM method, proposed algorithm.

Figure 3 presents the comparative results obtained by the proposed method and two other approaches of the literature: a pixel-based approach using GMM and a region-based method proposed by Sheikh and Shah [6]. We can notice that the results obtained by GMM are very noisy due to sudden illumination changes of the scene, while the detection accuracy of two region-based methods is still high.

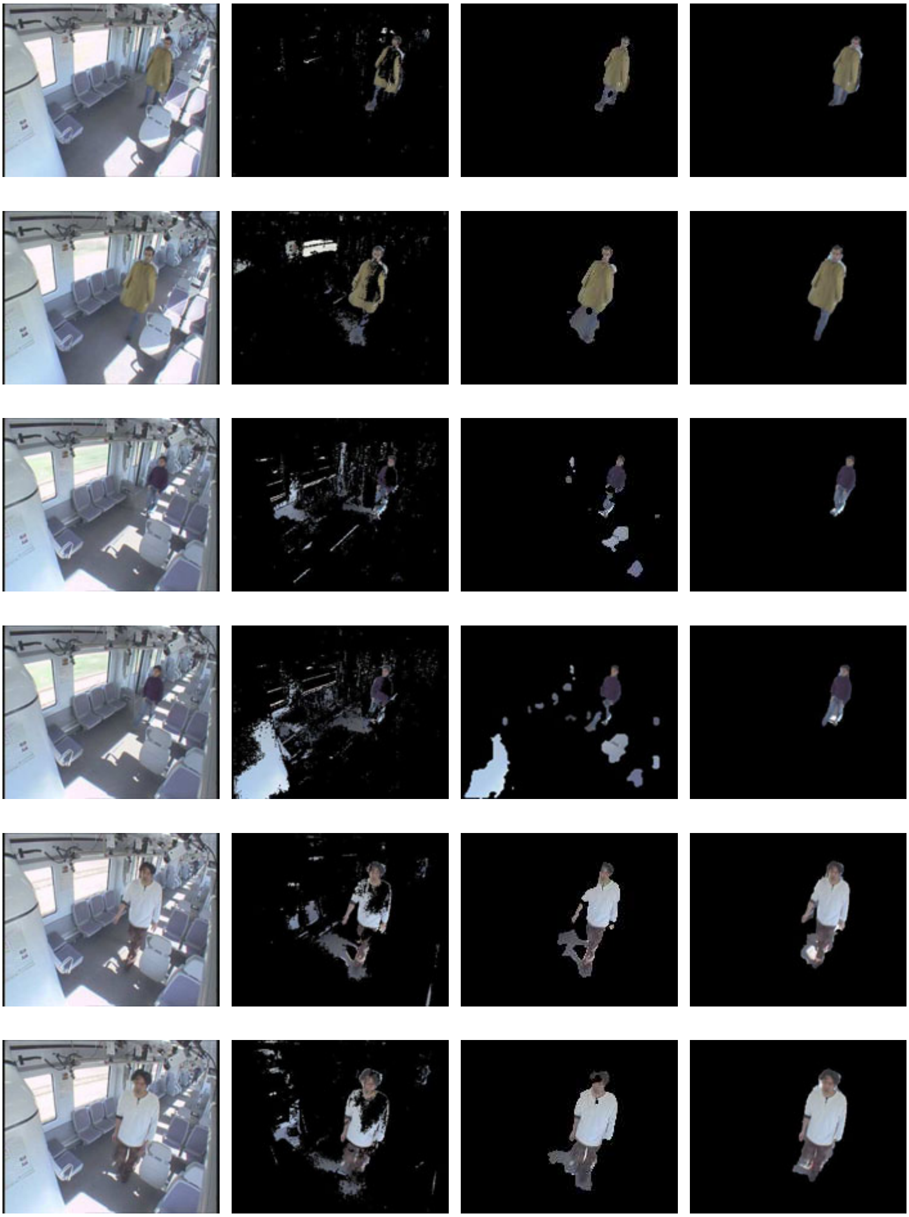


Fig. 3. Foreground object extraction results. Left to right: original images, GMM method, method proposed by Sheikh and Shah, our proposed algorithm.

In order to perform a quantitative analysis of the proposed approach, we have manually segmented 700 frames of a long sequence illustrated in Figure 3. The performances of the system are evaluated by using recall and precision measurements, where

$$\text{recall} = \frac{\text{number of true foreground pixels detected}}{\text{number of true foreground pixels}}$$

$$\text{precision} = \frac{\text{number of true foreground pixels detected}}{\text{number of foreground pixels detected}}$$

Table 1 presents the evaluation results in terms of recall and precision of three methods: the GMM algorithm with optimal parameters, the method proposed by Sheikh and Shah, and our proposed approach. In order to make a fair comparison, morphological operations are also used in the tests of standard GMM method. Clearly, the results demonstrate that the proposed approach obtains best performance in both terms of recall and precision.

Table 1. Comparative results in terms of recall and precision

	GMM	SS05	Proposed approach
Recall	0.54	0.91	0.95
Precision	0.73	0.91	0.94

Figure 4 shows the per-frame detection accuracy in terms of recall and precision. One can notice that our method is slightly more robust than the method proposed by Sheikh and Shah, and the extraction accuracies of the two region-based approaches are consistently higher than the standard GMM method.

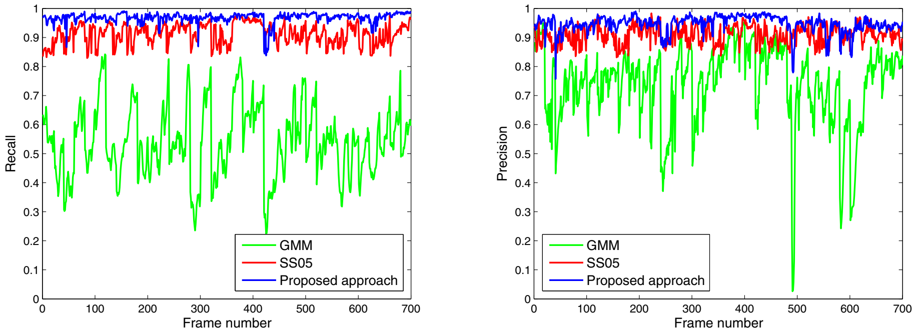


Fig. 4. Recall and precision curves obtained from the tested sequence

4 Conclusion

In this paper, we have presented a specific algorithm for foreground object extraction in complex scenes with non-stationary background. Several originalities are introduced to manage this difficult problem. A region-wise model of background and foreground is first proposed by using an adaptive mixture of Gaussians in a joint spatio-colorimetric feature space. A great robustness is introduced thanks to the simultaneous exploitation of background and foreground models. A dynamic decision framework, which is able to take advantages of both spatial and temporal coherency of object, is introduced for classifying background and foreground pixels. The proposed method was tested on a dataset coming from a real surveillance system including different sensors installed on board a moving train. The experimental results show that the proposed algorithm is robust in these real difficult scenarios.

In order to further improve the performance of the system and to reduce false detections caused by shadows, a normalized color space could be used instead of the RGB space. Moreover, several features (texture, edge,...) should be considered and integrated to the system to manage the cases where foreground and background colors are very similar.

References

1. Elhabian, S.Y., El-Sayed, K.M., Ahmed, S.H.: Moving object detection in spatial domain using background removal techniques - state-of-art. Recent Patents on Computer Science 1(1), 32–54 (2008)
2. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P.: Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 780–785 (1997)
3. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 246–252 (1999)
4. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000)
5. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Background modeling and subtraction by codebook construction. In: International Conference on Image Processing, vol. 5, pp. 3061–3064 (2004)
6. Sheikh, Y., Shah, M.: Bayesian modeling of dynamic scenes for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(11), 1778–1792 (2005)
7. Heikkila, M., Pietikainen, M.: A texture-based method for modeling the background and detecting moving objects. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4), 657–662 (2006)
8. Chen, Y.T., Chen, C.S., Huang, C.R., Hung, Y.P.: Efficient hierarchical method for background subtraction. Pattern Recognition 40(10), 2706–2715 (2007)
9. Dickinson, P., Hunter, A., Appiah, K.: A spatially distributed model for foreground segmentation. Image and Vision Computing 27(9), 1326–1335 (2009)
10. Nock, R., Nielsen, F.: Statistical region merging. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(11), 1452–1458 (2004)
11. Von Neumann, J., Burks, A.W.: Theory of Self-Reproducing Automata. University of Illinois Press Champaign, IL (1966)
12. http://www.multitel.be/boss