A new motion detection algorithm based on Sigma-Delta background estimation

Antoine Manzanera, Julien Richefeu

A new motion detection algorithm based on Σ-Δ background estimation

Antoine Manzanera, Julien C. Richefeu
ENSTA/LEI, 32 Bd Victor, F-75739, PARIS CEDEX 15, France

Abstract

Motion detection using a stationary camera can be done by estimating the static scene (background). In that purpose, we propose a new method based on a simple recursive non linear operator, the Σ-Δ filter. Used along with a spatiotemporal regularization algorithm, it allows robust, computationally efficient and accurate motion detection. To deal with complex scenes containing a wide range of motion models with very different time constants, we propose a generalization of the basic model to multiple Σ-Δ estimation.

Key words: Motion detection; Background estimation; Recursive filtering

1. Introduction

The detection of moving objects in an image sequence is a very important low-level task for many computer vision applications, such as video surveillance, traffic monitoring or sign language recognition. When the camera is stationary, a class of methods usually employed is *background subtraction*. The principle of these methods is to build a model of the static scene (i.e. without moving objects) called *background*, and then to compare every frame of the sequence to this background in order to discriminate the regions of unusual motion, called *foreground* (the moving objects).

Many algorithms have been developed for background subtraction: recent reviews and evaluations can be found in (Lee and Hedley, 2002) (Chalidabhongse et al., 2003) (Cheung and Kamath, 2004) (Piccardi, 2004). In this paper, we are more specifically interested in video surveillance systems with long autonomy. The difficulty in devising background subtraction algorithms in such context lies in the respect of several constraints:

- The system must keep working without human interaction for a long time, and then take into account gradual or sudden changes such as illumination variation or new static objects settling in the scene. This means that the background must be *temporally adaptive*.
- The system must be able to discard irrelevant motion such as waving bushes or flowing water. It should also be robust to slight oscillations of the camera. This means that there must be a *local* estimation for the *confidence* in the background value.
- The system must be real-time, compact and low-

Email address: antoine.manzanera@ensta.fr, julien.richefeu@ensta.fr (Antoine Manzanera, Julien C. Richefeu).

power, so the algorithms must not use much resource, in terms of computing power and memory.

The two first conditions imply that statistical measures on the temporal activity must be locally available in every pixel, and constantly updated. This excludes any basic approach like using a single model such as the previous frame or a temporal average for the background, and global thresholding for decision.

Some background estimation methods are based on the analysis of the histogram of the values taken by each pixel within a fixed number K of past frames. The mean, the median or the mode of the histogram can be chosen to set the background value, and the foreground can be discriminated by comparing the difference between the current frame and the background with the histogram variance. More sophisticated techniques are also based on the K past frames history: linear prediction (Toyoma et al., 1999), kernel density estimation (Elgammal et al., 2000) (Mittal and Paragios, 2004), or principal component analysis (Oliver et al., 2000). These methods require a great amount of memory, since K needs to be large (usually more than 50) for robustness purposes. So they are not compatible with our third condition.

Much more attractive for our requirements are the *recursive* methods, that do not keep in memory a histogram for each pixel, but rather a fixed number of estimates computed recursively. These estimates can be the mean and variance of a Gaussian distribution (Wren et al., 1997), different states of the background (e.g. its values and temporal derivatives) estimated by predictive filter (Karmann and von Brandt, 1990), or recursive estimation of the extremal values (Richefeu and Manzanera, 2004). But it is difficult to get robust estimates of the background with linear recursive framework, unless a multi-modal distribution (e.g. multiple Gaussian (Stauffer and Grimson, 2000) (Power and Schoonees, 2002)) is explicitly used, which is done at the price of an increasing complexity and memory requirement. Furthermore, these methods rely on parameters such as the learning rates used in the recursive linear filters, setting the relative weights of the background states and the new observations, whose tuning can be tricky, which makes difficult the fulfillment of the first condition stated above.

A recursive approximation of the temporal median was proposed in (McFarlane and Schofield, 1995) to compute the background. The interest of this method lies in the robustness provided by the non linearity compared to the linear recursive average, and in the very low computational cost. In this article, we investigate some nice properties of this operator, introducing the notion of Σ-Δ *estimation*, and using it to obtain a locally adaptive motion detection.

In Section 2, we present the basic Σ-Δ estimation method. The Σ-Δ filter is presented and used to compute two orders of temporal statistics for each pixel of the sequence providing a pixel-level decision framework. Then, in Section 3, we exploit the spatial correlation in these data using new hybrid linear/morphological operators, and use higher level processing to enhance and regularize the detection solution. Some results are presented, illustrating the robustness and accuracy of the method in the case of simple background (i.e. one single time-varying mode). For more complex scenes, we propose in Section 4 a generalization of the algorithm to multiple background estimation. Finally, conclusions and future works are presented in Section 5.

2. Σ-Δ estimation

Our first background estimate, whose computation is shown on Table 1(1), is the same as (McFarlane and Schofield, 1995), where I_t is the input sequence, and M_t the estimated background value. The sign function sgn is defined as $sgn(a) = -1$ if $a < 0$, $sgn(a) = 1$ if $a > 0$, and $sgn(a) = 0$ if $a = 0$. So, at every frame, the estimate is simply incremented by one if it is smaller than the sample, or decremented by one if it is greater than the sample. If I_t is a discrete random signal, the ratio between the number of indexes $\tau < t$ such that $I_\tau < M_t$, and the number of indexes $\tau < t$ such that $I_\tau > M_t$ converges in mean to 1. So M_t is an approximation of the median of I_t. But this filter has other interesting properties, relative to the change detection in time-varying signals. Indeed, we interpret this

background estimation as the simulation of a digital conversion of a time-varying analog signal using Σ-Δ modulation (A/D conversion using only comparison and elementary increment/decrement, hence the name Σ-Δ filter).

As the precision of the Σ-Δ modulation is limited to signals with absolute time-derivative less than unity, the modulation error is proportional to the variation rate of the signal, corresponding here to a motion likelihood measure of the pixels. We then use the absolute difference between I_t and M_t as our first differential estimate: the difference Δ_t (Table 1(2)).

Unlike (McFarlane and Schofield, 1995), we also use this filter to compute the time-variance of the pixels, representing their motion activity measure, used to decide whether the pixel is more likely "moving" or "stationary". Then, V_t (Table 1(3)) used in our method has the dimension of a temporal standard deviation. It is computed as a Σ-Δ filter of the difference sequence Δ_t. This provides a measure of *temporal activity* of the pixels. As we are interested in pixels whose variation rate is significantly over its temporal activity, we apply the Σ-Δ filter to the sequence of N times the non-zero differences.

Finally, the pixel-level detection is simply performed by comparing Δ_t and V_t (Table 1(4)).

Figure 1 displays an example of the evolution over time of the different values computed as above, for three particular pixels extracted from a country scene for a 500 frames sequence. The solid line represents the input image I_t. The dashed line corresponds to the Σ-Δ mean M_t. The impulses represents the difference Δ_t. Finally, the dotted line is the Σ-Δ variance V_t (using $N = 4$). The detection label D_t is not represented explicitly, but corresponds to the Boolean indicator of the condition "an impulse is over the dotted line".

The pixel used in Figure 1(1) is a pixel in a still zone, with flat temporal activity, such as a remote area of the static background (in our example, a sky lightly covered with slowly moving clouds). For such pixels, the high frequency variation corresponds to temporal noise due to the acquisition and digitization processes. The low frequency variations are due to illumination changes or slow motion of low contrast objects.

Initialization

for each pixel x:

$M_0(x) = I_0(x)$

For each frame t

for each pixel x:

$M_t(x) = M_{t-1}(x) + sgn(I_t(x) - M_{t-1}(x))$

(1)

For each frame t

for each pixel x:

$\Delta_t(x) = |M_t(x) - I_t(x)|$

(2)

Initialization

for each pixel x:

$V_0(x) = \Delta_0(x)$

For each frame t

for each pixel x such that $\Delta_t(x) \neq 0$:

$V_t(x) = V_{t-1}(x) + sgn(N \times \Delta_t(x) - V_{t-1}(x))$

(3)

For each frame t

for each pixel x:

if $\Delta_t(x) < V_t(x)$

then $D_t(x) = 0$

else $D_t(x) = 1$

(4)

Table 1
The Σ-Δ background estimation: (1) Computation of the Σ-Δ mean. (2) Computation of the difference between the image and the Σ-Δ mean (motion likelihood measure). (3) Computation of the Σ-Δ variance defined as the Σ-Δ mean of N times the non-zero differences. (4) Computation of the motion label by comparison between the difference and the variance.

The pixel used in Figure 1(2) is a pixel in a motion area, such as tracks or corridors (in our example, a country road with 2 vehicles passing away). In that case, the moving objects give rise to sharp changes that are not taken into account by the Σ-Δ mean, and then the difference signal shows a peak. Such peaks are discriminated thanks to the

comparison with the Σ-Δ variance.

The pixel used in Figure 1(3) is a pixel in a clutter area, i.e. a zone of physical changes due to intrinsic nature of the scene rather than moving objects. Examples of such areas are: trees moving with the wind or river (in our example, high grass in the foreground of the scene). In that case, the difference signal shows a repetition of peaks, and if these peaks are close enough from each other with respect to the delay induced by Σ-Δ modulation, then they will be taken into account in the Σ-Δ variance, in such a way that the difference will remain less that the variance.

The fundamental features of the Σ-Δ estimation, i.e. its non linearity and median convergence property come from the fact that the statistics M_t is always updated with a constant increment ± 1, not depending upon the difference between the current sample and the current mean $(I_t - M_{t-1})$. If the increment depended linearly on the difference, we would get $M_t = M_{t-1} + \alpha(I_t - M_{t-1}) = \alpha I_t + (1-\alpha)M_{t-1}$, that is, the classical recursive exponential filter (or moving average) used, typically, in the recursive Gaussian Fitting methods (Wren et al., 1997) (Stauffer and Grimson, 2000) (Power and Schoonees, 2002). In its simplest form, α is a real constant in the interval $]0, 1[$. Figure 2 (1)-(2) displays comparison between the Σ-Δ mean and the moving average, on a temporal sequence corresponding to the passage of a moving object. Note the difference between the unity slope of the Σ-Δ mean (Figure 2(1)) and the exponential decrease of the moving average (Figure 2(2)). In the more robust form of the Gaussian fitting estimation, α is no more a constant but is calculated from the probability of the current sample, i.e. $\alpha(t) = \mathcal{N}_{V_t}^{M_t}(I_t)$, where $\mathcal{N}_{\sigma^2}^{\mu}(x)$ is the normal density function of mean μ and variance σ^2. The variance V_t is estimated by the moving average of the squared differences $(I_t - M_t)^2$. Computation of the square and of the Gaussian probability of the sample make these methods sensitive to the numerical approximations, whereas the Σ-Δ estimation only uses exact integer computations. However, a nice feature of the Gaussian fitting methods is their extension to the estimation of multimodal Gaussian distributions (Stauffer and Grimson, 2000) (Power and Schoonees, 2002), that allows to deal with very complex scenes. In the same spirit, we shall present, in Section 4, an extension of the Σ-Δ estimation adapted to complex backgrounds.

Figure 2(3) also show the behavior another simple recursive estimation method, based on the forgetting morphological operators (Richefeu and Manzanera, 2004), because they will be used later in our algorithm, under the form of their spatial counterparts (See Section 3). The forgetting temporal dilation (resp. erosion) is defined by $M_t = \alpha I_t + (1-\alpha)\max(I_t, M_{t-1})$ (resp. $m_t = \alpha I_t + (1-\alpha)\min(I_t, m_{t-1})$).

Figure 3 (1)-(4) displays the result of the algorithm of Table 1 for one frame of an urban traffic sequence. The four images represent respectively I_t, M_t, V_t and D_t. It can be seen that the discrimination of "moving" pixels corresponds to the detection of temporally *salient* pixels with respect to the temporal activity. This allows to discard irrelevant (clutter) motion, but also to be less sensitive to sensor oscillations, as it is shown on Figure 3 (5) and (6): A uniform random oscillation of ± 1 pixels has been simulated on the same sequence. In this case, M_t converges to an approximation of the spatiotemporal median, and then V_t (Fig. 3(5)) emphasizes the regions of high contrast, thus increasing the local threshold in these regions, and avoiding the detection of the whole scene contour in D_t (Fig. 3(6)).

The only visible parameter of this method is N, the number used in the computation of the variance V_t. The range value of N is small (between 1 and 4), and usually a power of 2 is chosen for optimization purposes. Furthermore, it can be automatically adjusted with respect to a noise estimation. Such estimation can be performed by counting the isolated pixels in the detection result D_t, under the (classical) hypothesis that such detected pixels are only due to noise.

In fact, there is another parameter which is less obvious; it is the frequency of update of the Σ-Δ statistics. This frequency has a dimension of number of gray levels per second. It is then clear that it has to be adapted to (1) the dynamics of the image (number of gray levels), (2) the acquisition frequency (frame rate). We usually use the same frequency as the frame rate for 25 Hz sequences of

8-bits images, but it can be lowered to adjust to the size and velocity of the observed objects in the application. In Section 4, we will present a sophistication of the method based on a multi-frequencies Σ-Δ estimation, in order to better discriminate the foreground in the case of a scene presenting a wide range of different motions.

As suggested by Lacassagne in (Denoulet et al., 2005), the robustness of the Σ-Δ estimation can be further improved by updating M_t only when $\Delta_t < V_t$. In this article, we make the choice to inhibit locally the update only after the spatial processing (see Section 3).

The Σ-Δ background estimation provides a simple and efficient method to detect the significantly changing pixels in a static scene, with respect to a time constant depending on the number of gray levels, and on the frame rate. Nevertheless it is a pure temporal processing, which can only provide pixel-level detection. In the following section, we present some spatial processing, to enhance and regularize the detection result.

3. Spatiotemporal processing

Recently, we have presented a Markovian modeling to perform a spatiotemporal regularization of the pixel-level Σ-Δ detection. It was an adaptation of the iterative algorithm presented in (Caplier et al., 1996) and (Lacassagne et al., 1999), using the pixel-level detection D_t as initialization, and the Σ-Δ difference Δ_t and variance V_t, as a couple of observation fields used in the design of the energy. Details can be found in (Manzanera and Richefeu, 2004).

We present here another regularization strategy. The spatiotemporal processing that we propose in this section has a threefold purpose:

- eliminate the *non significant* pixels from the detection (noise, false detection), and enhance the segmentation of the moving objects.
- reduce the *ghost effect*, that produces false detection at the loci from where a moving object leaves after a long stay.
- reduce the *aperture effect*, that causes a poor detection for the objects whose projected motion is weak, e.g. radially moving objects.

3.1. *Common edges hybrid reconstruction*

The first part of the spatial processing is composed of gray level operations. The inputs are: (1) the original image I_t, and (2) the Σ-Δ difference image Δ_t. The purpose of this module is to eliminate the ghost objects in Δ_t by discriminating them within I_t. The actual operation we perform is the following one:

$$\Delta'_t = HRec_{\alpha}^{\Delta_t}(Min(||\nabla(I_t)||, ||\nabla(\Delta_t)||)) \quad (1)$$

Roughly speaking, this means the "reconstruction" (we shall make this word explicit later) within Δ_t, of the image of the minimum between the gradient module of Δ_t and the gradient module of I_t. The semantics of this formula is: "Δ'_t is made of the components of Δ_t that are also in I_t". Let us now detail the actual computation.

The gradient modules of Δ_t and I_t are computed by estimating the first derivative components with convolutions with Sobel masks, and then computing the Euclidean norm of the vector. We then compute the minimum image Min, defined for every pixel x by $Min(I_1, I_2)(x) = \min(I_1(x), I_2(x))$. This acts like a logical conjunction, retaining only the edges that belongs both to an object of Δ_t, and of I_t.

In order to recover the whole object in Δ_t, and not only its edges, we perform a "reconstruction" of the common edges $Min(||\nabla(I_t)||, ||\nabla(\Delta_t)||)$ within Δ_t.

The first idea should be to use the classical geodesic reconstruction $Rec^Y(X)$, defined by the relaxation of the geodesic dilation $\delta_B^Y(X) = Min(\delta_B(X), Y)$ (δ is the morphological dilation, B the elementary structuring element defining the topology, Y the reference image, X the marker image). In fact the geodesic reconstruction is not adapted, because the sole connection criterion is not robust enough, and in most cases, the object and its ghost are both reconstructed.

We rather use the *hybrid reconstruction* operator, based on the forgetting morphological operators that we introduced in (Richefeu and Manzanera, 2004). First we define the hybrid dilation as

the spatial version of the forgetting dilation, computed by the causal sequence:

$$HDil_\alpha(I)^{(0)}(x,y) = \alpha I(x,y) + (1-\alpha)\max(I(x,y), HDil_\alpha(I)^{(0)}(x-1,y)) \quad (2)$$

followed by the anti-causal sequence:

$$HDil_\alpha(I)^{(1)}(x,y) = \alpha HDil_\alpha(I)^{(0)}(x,y) + (1-\alpha)\max(HDil_\alpha(I)^{(0)}(x,y), HDil_\alpha(I)^{(1)}(x+1,y)) \quad (3)$$

then the causal sequence vertically:

$$HDil_\alpha(I)^{(2)}(x,y) = \alpha HDil_\alpha(I)^{(1)}(x,y) + (1-\alpha)\max(HDil_\alpha(I)^{(1)}(x,y), HDil_\alpha(I)^{(2)}(x,y-1)) \quad (4)$$

and finally:

$$HDil_\alpha(I)(x,y) = \alpha HDil_\alpha(I)^{(2)}(x,y) + (1-\alpha)\max(HDil_\alpha(I)^{(2)}(x,y), HDil_\alpha(I)(x,y+1)) \quad (5)$$

Here $1/\alpha$ has the dimension of a spatial radius, and thus the parameter α replaces the structuring element.

The hybrid reconstruction is based on the same scheme, using the sequence:

$$HRec_\alpha^J(I)^{(0)}(x,y) = \min(J(x,y), \alpha I(x,y) + (1-\alpha)\max(I(x,y), HRec_\alpha^J(I)^{(0)}(x-1,y))) \quad (6)$$

and so on.

The behavior of the hybrid reconstruction can be seen on Figure 4, where the objects (cars, pedestrian) move after leaving a halt 20 frames before. The advantage of the hybrid reconstruction in this application is to "forget" gradually (exponentially to be precise) the marker, which acts like a confidence function, instead of being strictly based on the connectivity.

It can be seen that the success of this common edges marking step depends on the contrasts of the background itself, (see Figure 4(7)-(10)). But we have focused here on extreme cases, the objects being completely still and encrusted in the background at the beginning. Furthermore, as we will see at the end of this section, the relevance feedback will also reduce significantly the ghost effect.

3.2. *Binary spatiotemporal morphology*

After computing the hybrid reconstruction, we perform the adaptive thresholding on Δ'_t (instead of Δ_t in the purely temporal version):

$$\text{If } \Delta'_t > V_t \text{ then } D_t = 1 \text{ else } D_t = 0 \quad (7)$$

We then eliminate the small connected components on D_t using the *opening by reconstruction*:

$$L_t^{(0)} = Rec^{D_t}(\varepsilon_{B_\lambda}(D_t)) \quad (8)$$

where ε is the morphological erosion, and B_λ the structuring element, is a ball of radius λ.

Finally, we perform a *temporal confirmation* by computing another reconstruction:

$$L_t = Rec^{L_t^{(0)}}(L_{t-1}^{(0)}) \quad (9)$$

L_t represents the final label, the result of the detection. Its semantics is: the objects "bigger than" λ that appear on 2 consecutive frames. It is illustrated on Figure 5, using $\lambda = 1$, on a sequence showing two people walking.

3.3. *Relevance feedback*

As indicated in Section 2, the whole detection is made more robust if the background is not modified for the pixels of moving objects. We adopt this strategy by updating Σ-Δ mean $M_t(x)$ only where $L_t(x) = 0$. This implies a reordering of the algorithm shown in Table 1. The complete algorithm is then displayed in Table 2, discarding the pixel index x.

The relevance feedback effect can be seen on Figure 6, showing a detail of a scene with 2 cars driving after a stop at a red light. This strategy improves the detection result by delaying the contribution of the moving objects to the background. This reduces significantly both the ghost effect and the aperture effect (radial movements).

Signed difference

$S_t = I_t - M_{t-1}$

Variance updating

if $|S_t| \neq 0$:

$V_t = V_{t-1} + sgn(N \times |S_t| - V_{t-1})$

Common edges hybrid reconstruction

$\Delta'_t = HRec_{\alpha}^{|S_t|}(Min(||\nabla(I_t)||, ||\nabla(|S_t|)||))$

Temporal detection

if $\Delta'_t < V_t$

then $D_t = 0$

else $D_t = 1$

Spatiotemporal binary processing

$L_t^{(0)} = Rec^{D_t}(\varepsilon_{B\lambda}(D_t))$

$L_t = Rec^{L_t^{(0)}}(L_{t-1}^{(0)})$

Mean updating with relevance feedback

if $L_t = 0$:

$M_t = M_{t-1} + sgn(S_t)$

Table 2
The complete Σ-Δ detection algorithm with relevance feedback.

As indicated in (Denoulet et al., 2005), it is safer, at this low level of processing, to apply the relevance feedback only to the mean M_t instead than the two estimates M_t and V_t, to avoid false detection objects to settle in the background. Relevance feedback can be used on both M_t and V_t if a high level of confidence is reached in the detected objects. This can be achieved by using further processing (e.g. kinematic filtering), that is beyond the scope of this article.

Figure 7 shows the results of the full algorithm on an indoor sequence (the "Hall" sequence). The results illustrate the effect of the spatial processing and relevance feedback on the reduction of the aperture effect: the two persons are well detected although they are moving radially with respect to the camera.

4. Multiple background Σ-Δ estimation

The spatiotemporal processing and relevance feedback, presented in the previous section, allows a visible enhancement of the robustness, in the case of slowing down, stopping or radially moving objects. Nevertheless, the Σ-Δ estimator is characterized by a time constant: its updating period, which has a dimension of number of gray levels per second. This induces a limitation of the basic approach in the adaptation capability to certain complex scenes, typically in the case of scenes permanently crossed by lots of objects of very different sizes and velocities. We propose in this section a generalization framework of the Σ-Δ estimation to multiple backgrounds.

The principle is to compute instead of one single background M_t, a set of K backgrounds $\{m_t^i\}_{1 \leq i \leq K}$. Each background m_t^i is characterized by its updating period α_i and by its phase ϕ_i. A set of K variances v_t^i is also computed as the Σ-Δ mean of the differences between I_t and m_t^i. The background/foreground decision is then made by comparing the sample to every background, which is attached a confidence value that is (1) proportional to α_i and (2) inversely proportional to v_t^i.

Table 3 shows an example of computation of multi-frequencies background using K different periods $\alpha_1 < \ldots < \alpha_K$. The phases are discarded in this example. The Σ-Δ means m_t^i can be computed recursively: $m_t^i = m_{t_1}^i + sgn(m_t^{i-1} - m_{t_1}^i)$, with the convention $m_t^0 = I_t$. In this example, we compute explicitly a "best confidence" background M_t, as shown in the last line of Table 3. The principle of the confidence coefficients attached to every background m_t^i is to give more weight to long-term backgrounds, in order to favor the most frequent value over long periods, and at the same time, the weights are lowered accordingly to the variance, in order to allow adaptation to a changing background.

Figure 8 shows an application of the multi-periods Σ-Δ estimation on a complex scene. This corresponds to a sequence taken in a very frequented part of the Jardin du Luxembourg. The scene is never empty, with lots of people stopping and remaining more or less still for different pe-

Multiple mean updating

For each frame t,

for each i, $i \in [0, K]$,

if t is a multiple of α_i:

$m_t^i = m_{t_1}^i + sgn(m_t^{i-1} - m_{t_1}^i)$

Multiple variance updating

For each frame t,

for each i, $i \in [0, K]$,

$\delta_t^i = |I_t - m_t^i|$

if $\delta_t^i \neq 0$:

$v_t^i = v_{t_1}^i + sgn(v_{t-1}^i - N \times \delta_t^i)$

Mean computing

For each frame t,

$$M_t = \frac{\sum_{i \in [0,K]} \frac{\alpha_i m_t^i}{v_t^i}}{\sum_{i \in [0,K]} \frac{\alpha_i}{v_t^i}}$$

Table 3
The multi-periods background Σ-Δ estimation.

riods. For this sequence, we have taken $K = 3$, $\alpha_1 = 1$, $\alpha_2 = 8$, $\alpha_3 = 16$. This framework clearly enhances the robustness of the detection with respect to the range of motion models, while preserving a good adaptation capability to changing conditions. See for example on Figure 8: At frame (a), best confidence is globally given to the long-term mean. At frame (b), the camera has been shifted just before, and so, in the upper part of the image, best confidence is given to the short-term mean, while in the lower part of the image, the confidence remains higher for the long-term, which prevents the two stopped people to appear in the final background.

5. Conclusions

We have presented a new algorithm allowing a robust and accurate detection of moving objects for a small cost in memory consumption and computational complexity. We have emphasized the nice properties of the Σ-Δ filter for the detection of salient features in time-varying signal, showing that the interest of such filter goes well beyond its temporal median convergence property.

We have proposed a new spatiotemporal regularization strategy, using an original hybrid reconstruction method and spatiotemporal binary morphology, to exploit the spatial correlation and increase the confidence of the Σ-Δ detection.

We have presented a generalization framework for multiple Σ-Δ estimation allowing to deal with complex scenes by combining Σ-Δ estimates with different frequencies or phases.

Because it only relies on pixel-wise or spatially limited interactions, the main part of this algorithm (everything excepted the common edge hybrid reconstruction module, which is a recursive scan algorithm) is suited to a *massively parallel* implementation. We have realized an implementation of the algorithm (temporal processing plus spatiotemporal morphology) on a *programmable artificial retina* (Komuro et al., 2003), which is a fine-grained parallel machine with optical input. The algorithm is indeed well adapted to the architecture, which consists in a mesh of tiny processors with limited memory and computation power. For a 200x200 retina array running at 25 Mhz, using 8 bits per pixels: the computation performance is 2.25 ms per frame, of which only 0.75 ms for the sole computation, and the rest for the acquisition. (Denoulet et al., 2005) have also made an implementation of the Σ-Δ background estimation associated with Markovian relaxation on the *Associative Mesh of Orsay* (Mérigot, 1997) which is a massively parallel asynchronous machine with programmable topology.

Although we have seen how some complex scenes can be addressed with our approach, several limitations remain, for example wide amplitude periodical motion (e.g sea surge), or backgrounds with several equiprobable modes. We are investigating the possibilities to go beyond these limits by using estimates with different periods *and* phases, in order to get a richer quantitative estimation of the motion activity. Automatically adapting the different frequencies and phases to the observed signal should be the next challenge. We are also working on increasing the robustness of the background estimation, and then of the whole detection, by using

confidence feedback provided by higher level processing.

Acknowledgments

The authors wish to thank Lionel Lacassagne, who proposed improvements on the original algorithm, and provided most of the sequences used in this article.

References

Caplier, A., Dumontier, C., Luthon, F., Coulon, P., 1996. Mrf based motion detection algorithm image processing board implementation. Traitement du signal (in french).

Chalidabhongse, T. H., Kim, K., Harwood, D., Davis, L., 2003. A perturbation method for evaluating background subtraction algorithms. In: Proc. Joint IEEE Int. Work. on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Nice - France.

Cheung, S.-C., Kamath, C., 2004. Robust techniques for background subtraction in urban traffic video. In: Proc. SPIE Video Com. and Image Proc. San Jose - CA.

Denoulet, J., Mostafaoui, G., Lacassagne, L., Mérigot, A., 2005. Implementing motion markov detection on general purpose processor and associative mesh. In: Proc. CAMP'05.

Elgammal, A., Harwood, D., Davis, L., 2000. Nonparametric Model for Background Subtraction. In: Proc. IEEE ECCV. Dublin - Ireland.

Karmann, K.-P., von Brandt, A., 1990. Time-Varying Image Processing and Moving Object Recognition. Elsevier, Ch. Moving Object Recognition Using an Adaptive Background Memory.

Komuro, T., Ishii, I., Ishikawa, M., Yoshida, A., 2003. A digital vision chip specialized for high-speed target tracking. IEEE Trans. on Electron Devices.

Lacassagne, L., Milgram, M., Garda, P., 1999. Motion detection, labeling, data association and tracking in real-time on risc computer. In: Proc. IEEE ICIAP. pp. 520–525.

Lee, B., Hedley, M., 2002. Background estimation for video surveillance. In: Proc. IVCNZ'02. pp. 315–320.

Manzanera, A., Richefeu, J., Dec. 2004. A robust and computationally efficient motion detection algorithm based on Σ-Δ background estimation. In: Proc. ICVGIP'04. pp. 46–51.

McFarlane, N., Schofield, C., 1995. Segmentation and tracking of piglets in images. Machine Vision and Applications 8, 187–193.

Mérigot, A., 1997. Associative nets model: a graph based parallel computing model. IEEE trans. on Computers 46(5), 558–571.

Mittal, A., Paragios, N., 2004. Motion-based background subtraction using adaptive kernel density estimation. In: Proc. IEEE CVPR.

Oliver, N., Rosario, B., Pentland, A., 2000. A bayesian computer vision system for modeling human interactions. IEEE Trans. on PAMI.

Piccardi, M., Oct. 2004. Background subtraction techniques: a review. In: Proc. of IEEE SMC/ICSMC.

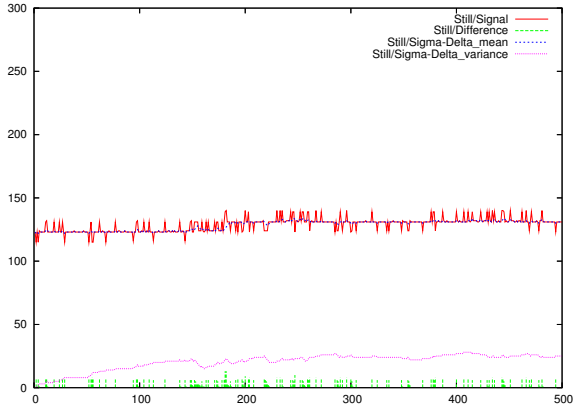
Power, P., Schoonees, J., Nov. 2002. Understanding background mixture models for foreground segmentation. In: Proc. IVCNZ'02. pp. 267–271.

Richefeu, J., Manzanera, A., Oct. 2004. A new hybrid differential filter for motion detection. In: Proc. ICCVG'04.

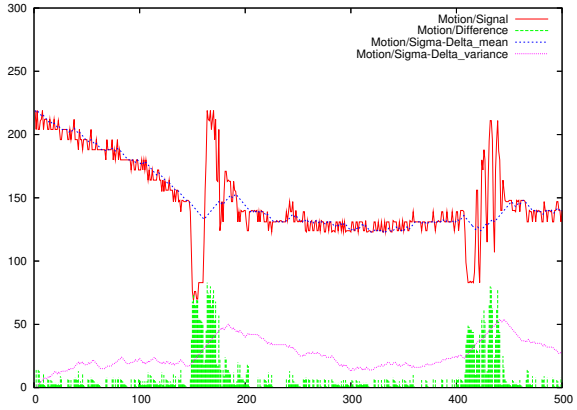
Stauffer, C., Grimson, E., 2000. Learning patterns of activity using real-time tracking. IEEE Trans. on PAMI.

Toyoma, K., Krumm, J., Brumitt, B., Meyers, B., 1999. Wallflower: Principles and Practice of Background Maintenance. In: Proc. IEEE ICCV. Kerkyra - Greece, pp. 255–261.

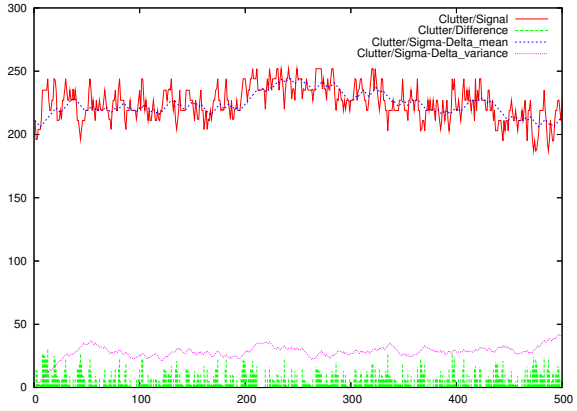
Wren, C., Azarbayejani, A., Darrell, T., Pentland, A., 1997. Pfinder: Real-time tracking of the human body. IEEE Trans. on PAMI.



(1)

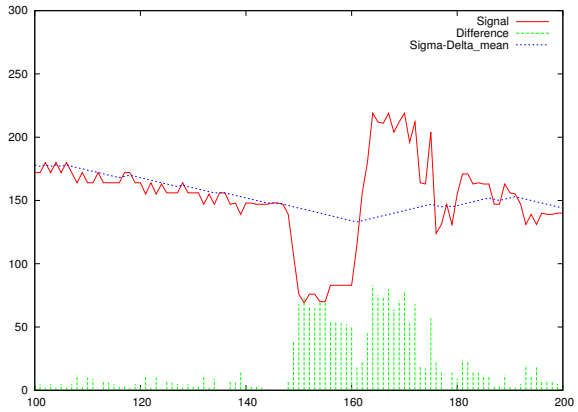


(2)

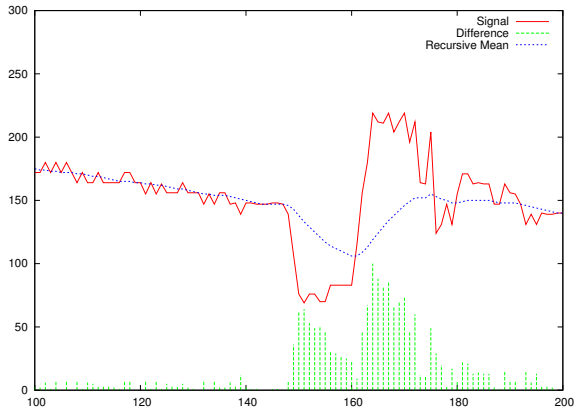


(3)

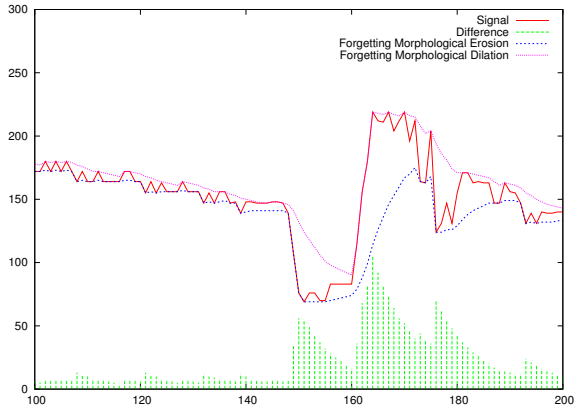
Fig. 1. Temporal variation of a pixel value and the corresponding Σ-Δ estimation, for pixels taken in three different areas (1) still area, (2) motion area, (3) clutter area.



(1)



(2)



(3)

Fig. 2. Comparison between (1) Σ-Δ estimation (2) Moving average ($\alpha = 1/16$). (3) Forgetting morphological operators ($\alpha = 1/8$).

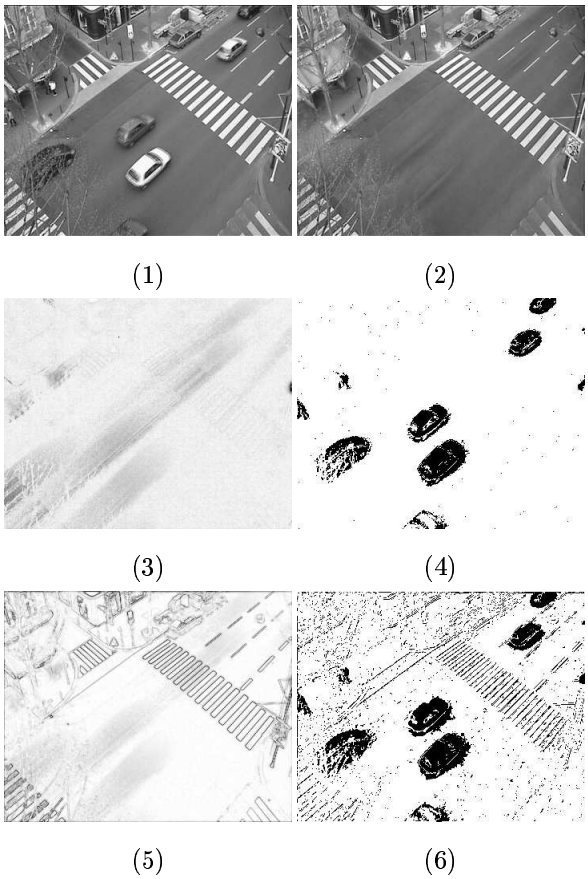


Fig. 3. Result of the Σ-Δ detection on a traffic sequence. (1) I_t (2) M_t. (3) V_t (displayed with reverse video and normalized histogram). (4) D_t (N=2). (5) V_t with simulated oscillations of the camera (displayed in reverse video and normalized histogram). (6) D_t for the oscillating camera (N=2).

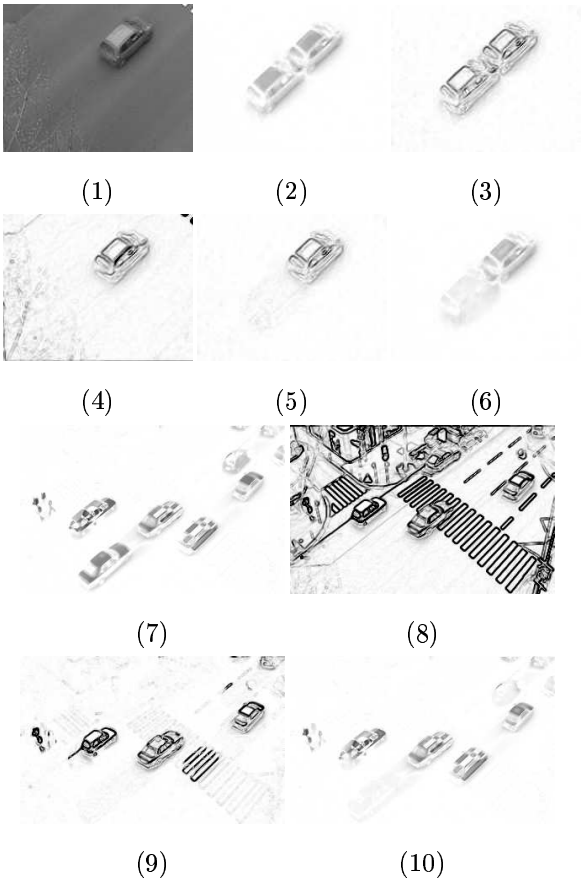


Fig. 4. Ghost busting by hybrid reconstruction of the common edges ($\alpha = 1/8$). First example (1 car): (1) I_t (2) Δ_t (3) $||\nabla(\Delta_t)||$ (4) $||\nabla(I_t)||$ (5) $Min(||\nabla(\Delta_t)||, ||\nabla(I_t)||)$ (6) Δ'_t. Second example (5 cars, 1 pedestrian): (7) Δ_t (8) $||\nabla(I_t)||$ (9) $Min(||\nabla(\Delta_t)||, ||\nabla(I_t)||)$ (10) Δ'_t.

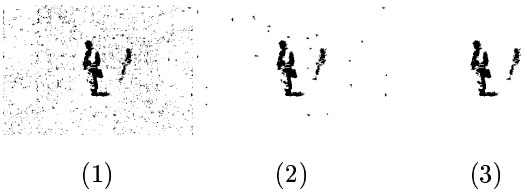
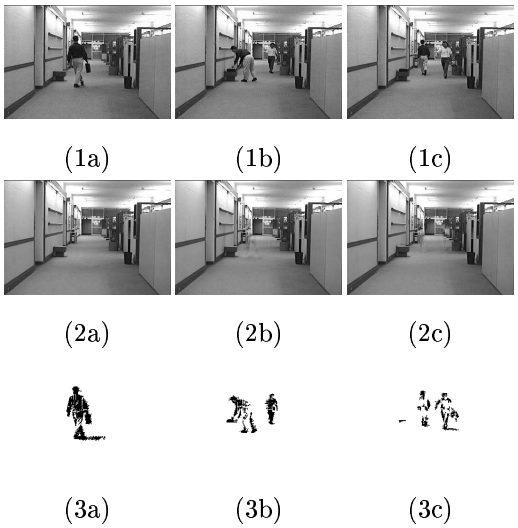


Fig. 5. Binary spatiotemporal morphology. (1) D_t (2) $L_t^{(0)}$ (3) L_t

(1) (2) (3)

Fig. 6. Relevance feedback. (1) Original image (2) Σ-Δ Background (3) Background with relevance feedback

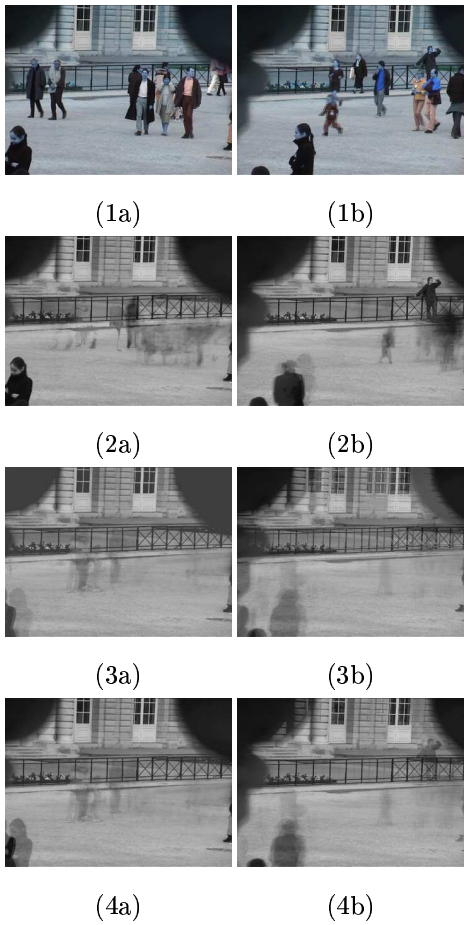


(1a) (1b) (1c)

(2a) (2b) (2c)

(3a) (3b) (3c)

Fig. 7. Result of the Σ-Δ detection with spatial processing on an indoor (Hall) sequence, shown at 3 frames: (a) t=60 (b) t=130 (c) t=200. (1) Original I_t (2) Σ-Δ mean M_t. (3) Detection label L_t.



(1a) (1b)

(2a) (2b)

(3a) (3b)

(4a) (4b)

Fig. 8. Multi-periods background Σ-Δ estimation computed on a complex scene (Luxembourg sequence), shown at two instants (a) t = 1136 (b) t = 2896. (1) Original I_t. (2) Short-term background m_t^1 ($\alpha_1 = 1$) (3) Long-term background m_t^3 ($\alpha_3 = 16$) (4) Best confidence background M_t.