Robust Motion Detection with Temporal Decomposition and Statistical Regularization

Jean-Michel Létang, Patrick Bouthemy, Véronique Rebuffel

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Robust motion detection with temporal decomposition and statistical regularization

Jean-Michel LÉTANG, Patrick BOUTHEMY &
Véronique REBUFFEL

N 2717

Octobre 1995

PROGRAMME 4

ISSN 0249-6399

Robust motion detection with temporal decomposition and statistical regularization

Jean-Michel LÉTANG*, Patrick BOUTHEMY** &
Véronique REBUFFEL***

Programme 4 — Robotique, image et vision
Projet Temis

Rapport de recherche n°2717 — Octobre 1995 — 33 pages

Abstract: This paper deals with the detection of moving objects. We have defined a method able to cope with perturbations frequently encountered during acquisition of outdoor image sequences: camera not perfectly stationary, illumination modifications, occlusions, ... Temporal integration and statistical regularization are the two main features of the method. A temporal multiscale decomposition allows us to detect and to characterize various dynamical behaviours of the elements present in the scene. A tracking module provides a prediction map, which gives a confidence level for presence of motion at a given pixel. A statistical regularization framework, based on Markov random field models, supplies a formal way to combine these different sets of computed information, while exploiting a priori knowledge on the primitives to be determined. A calibration technique based on so-called qualitative boxes is used to estimate model parameters. Several experiments with real image sequences depicting various complex situations have validated the approach.

Key-words: Image sequence, motion detection, multiresolution, temporal integration, statistical regularization, Markov models, parameter estimation, robustness.

(Résumé : tsvp)

*Jean-Michel.Letang@imag.fr (Unité de recherche Rhône-Alpes)
**Patrick.Bouthemy@irisa.fr
***LETI (CEA-DTA), CEN/G, 17 rue des martyrs, 38054 Grenoble Cedex 9, France, vero@dsys.ceng.cea.fr

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 99 84 71 00 – Télécopie : (33) 99 84 71 71

Détection de mouvement robuste basée sur une décomposition temporelle et une régularisation statistique

Résumé : Ce rapport traite de la détection d'objets mobiles en imagerie monoculaire. La première partie expose la problématique de l'analyse du mouvement dans des applications réelles. Nous proposons une méthodologie robuste vis à vis des perturbations fréquemment rencontrées lors de l'acquisition de scènes en extérieur. Nous dégageons trois directions de recherches qui mettent en relief l'importance de l'axe temporel, dimension privilégiée en analyse du mouvement. Dans une première partie, la séquence d'images est assimilée à un faisceau de signaux temporels. La décomposition multi-échelle temporelle mise en œuvre permet de caractériser les différents comportements dynamiques présents dans la scène à un instant donné. Un deuxième module intègre l'information de mouvement. Cette trajectographie élémentaire des objets mobiles permet d'obtenir une carte de prédiction temporelle, donnant un indice de confiance sur la présence d'un mouvement. Les interactions entre ces deux types de données sont exprimées au sein d'une régularisation statistique. La modélisation par champs de Markov fournit un cadre formel pour traduire des connaissances a priori sur les primitives à évaluer. Une méthode de calibrage par boîtes qualitatives est présentée pour estimer les paramètres de ce modèle. Notre approche se ramène à des calculs simples et conduit à une algorithmie relativement rapide, que nous évaluons en dernière partie sur des séquences variées typiques.

Mots-clé : Séquence d'images, détection de mouvement, multirésolution, intégration temporelle, régularisation statistique, modèles markoviens, estimation de paramètres, robustesse.

1 Introduction

Detection of moving objects in a scene is an issue of key interest in many applications. For indoor scenes, visual features (contrasted primitives such as corners, edges, ...) can be used as landmarks for motion analysis. Man-made environments which contain such special configurations are exploited in autonomous robot navigation. If illumination conditions can be controlled, it is easier to ensure satisfactory performance of the dynamic vision process. For outdoor scenes, in case of traffic surveillance for example, many approaches rely on hypotheses such as the use of a strictly static camera, or strong a priori knowledge on the objects, e.g., by introducing 3D models of vehicles. When coping with perturbations (camera instability, illumination variation, lack of distinguishable features, ...), in fact commonly faced during acquisition, deficiencies then come out.

We are concerned with motion detection in case of a stationary camera but, with emphasis on robustness to perturbations likely to occur in a realistic application context. Let us briefly quote several items that may significantly affect usual approaches:

- Sensor noise (quantization error, temporal aliasing, ...).
- No reliable features to detect and track moving objects.
- In case of a static camera, wind or camera handling may cause small oscillations. This induces an apparent small motion for contrasted contours in the image. Moreover, the viewing system may be placed in a device or a system (e.g., a ship) which may be not perfectly static or very slowly moving with respect to the observed mobile objects.
- Projection of moving objects can vary a lot according to their distance and attitude w.r.t. the camera. Apparent size can be very small. Transversal speed, that means velocity component parallel to the image plane, can be low. Moreover, shape of the projections of the moving objects can be affected by the surrounding environment (occlusions, ...).
- Illumination seldom remains constant in outdoor scenes: moving clouds may cause local or global lighting modifications for instance. Illumination change can also modify some of the camera intrinsic parameters if automatic setting is used, leading to global change in the image intensity distribution.

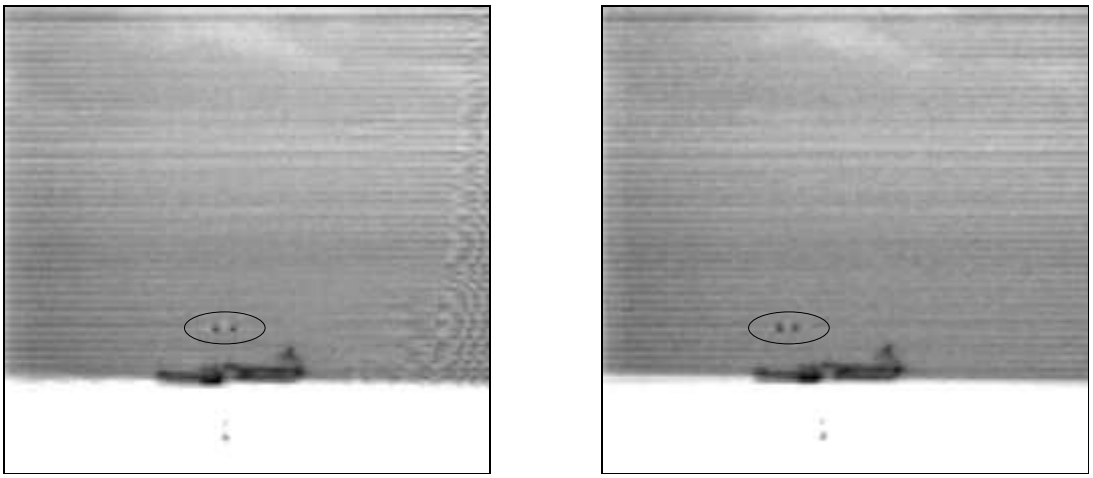

Figure 1: "Horizon" sequence. The moving plane is encircled. Images 0 (left) and 115 (right).

As an illustration, the scene shown in Figure 1 was acquired by a linear array infrared sensor. Dark spots corresponds to hot areas. The only moving object (circled in images 0 and 115) is a plane approaching above the ocean. In Figure 2, we can notice the intensity change of the pixel belonging to the plane trajectory due to the passing of both engine turbines. Two stationary boats overlap one another close to horizon line. A buoy floats in the foreground. Since the plane is coming towards the camera, its apparent motion magnitude is quite small: around 0.15 pixel per frame. Note that the acquisition system is not strictly static – camera is moving around its refrence position – leading to strong perturbations in the areas where spatial intensity gradient is high (see Figure 2). This scene will be referred to as the "horizon" sequence in the remainder of the paper.

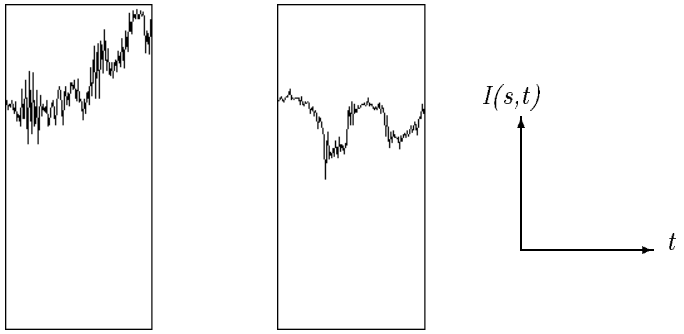

Figure 2: "Horizon" sequence. Intensity profile vs. time at a given pixel position s close to the horizon line (left), and for a pixel belonging to the plane trajectory (right).

Figure 3 shows a rural environment with two moving objects. These images were acquired by a infrared sensor (matrix type). Camera was held by hand. Both vehicles move from right to left. The vehicle in the middle is never completely visible, most of it is hidden by surrounding environment: several non-connected parts of this vehicle can be seen through the bushes. The vehicle on the right is more clearly visible, but its lower part has the same intensity as the ground. Two representative pixels, whose intensity profiles are plotted over time in Figure 4, illustrate that perturbations due to camera handling can almost be as important as those induced by the real motion. This scene is referred to as the "bush" sequence in the remainder of the paper.

Figure 3: "Bush" sequence. Vehicles are encircled. Images 0 (left) and 92 (right).

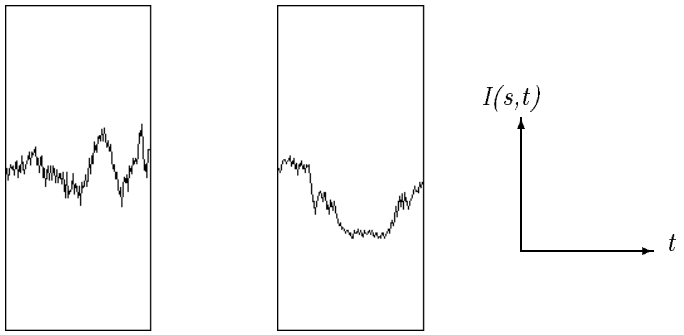

Figure 4: "Bush" sequence. Intensity profile vs. time at a given pixel location s in the foreground bush area (left), and for a pixel belonging to the right vehicle trajectory (right).

The development of a robust approach is then very important as we want to be able to deal with real outdoor sequences, without deteriorating the algorithm performance in the presence of various perturbations. Solving this problem could rely on a motion estimation step, but this approach would suffer from several shortcomings. First, computing optic flow fields is time consuming. The main point in fact is that the optic flow accuracy that we could obtain, would certainly not be high enough to perform a segmentation allowing us to locate small objects in motion. Therefore, it is preferable to adopt a motion detection approach, which is a simpler processing, provided we can adjust it to fulfil the robustness requirements. An overview of our approach is given in Figure 5, which clearly shows the three main modules of the method.

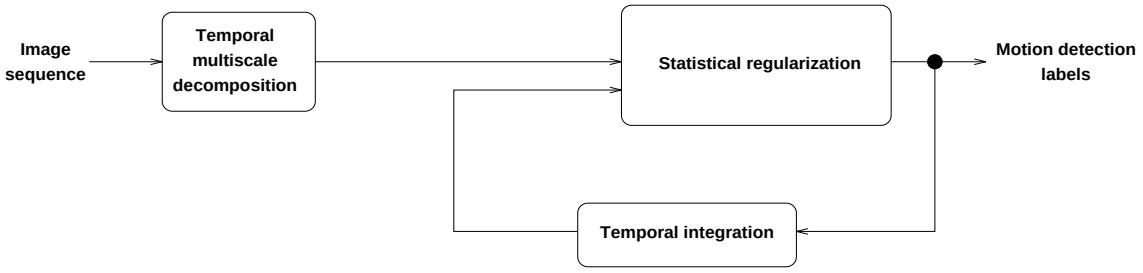

Figure 5: Overview of the motion detection method

The image sequence is considered as a set of 1D temporal signals, and a temporal multiscale decomposition allows us to detect and to characterize various dynamical behaviours of the elements present in the scene. A simple tracking module provides a prediction map, which gives a confidence level for presence of motion at a given pixel. A statistical regularization framework, based on Markov random field (MRF) models, supplies a formal way to combine these different sets of computed information, while exploiting a priori knowledge on the primitives to be determined. Indeed, in recent motion detection approaches, improvements were often gained by considering temporal integration [25, 34, 11]. On the other hand, fusion of various information is also beneficial, and it can be easily achieved using MRF. Let us give two examples dealing with other motion analysis issues. Heitz and Bouthemy [30] have combined feature-based measurements and spatio-temporal gradients to compute optical flow. Black [9] has incorporated spatial segmentation and temporal coherence constraints to perform motion estimation and segmentation.

The remainder of this paper is organized as follows. Section 2 comprises a brief review of related work concerning detection of moving objects. Section 3 describes a temporal multiscale decomposition and the likelihood ratio tests allowing us to get temporal change maps at different scales. A complementary tracking scheme is presented in Section 4. Section 5 shows how this motion detection issue is stated and solved within a statistical labelling framework based on MRF models. We address in Section 6 the model parameter estimation problem, using a qualitative "calibration" approach [3]. Experiments carried out with real image sequences are reported in Section 7. Section 8 contains concluding remarks.

2 Related work

Detection of moving objects, which is one fundamental goal in motion analysis, can simply involve temporal change detection if camera is static, but it can also require a complete motion segmentation step if the acquisition system itself is moving.

Hsu, Nagel and Rekers [31] have made use of statistical likelihood tests for intensity change detection between two consecutive frames. This method is more efficient when displacements are large enough or object projections are sufficiently textured. Since it only uses short range temporal information (as any two-frame method), it may have difficulty to balance robustness to noise and completeness of the detection. This test has often been used as part of motion detection schemes. It has also been exploited in some way in an illumination independent technique [45].

In order to guarantee robust detection results in the presence of noise in video sequences, Aach et al. have derived a MRF-based regularization of the temporal change maps, [1]. Furthermore, Bouthemy and Lalande [13] have developed an approach to formulate the motion detection problem (which represents a step further beyond change detection) using local spatio-temporal contextual information. This approach, based on statistical regularization, can handle objects of different size and different motion, but it only takes into account frame to frame intensity differences.

As far as spatio-temporal filters are concerned, if no a priori information is available about velocity magnitude and direction, the search space is usually very large. Fleet and Jepson [27] have designed Gabor filters, Liou and Jain [38] have built spatio-temporal gradient detectors.

Donohoe et al. [25] have introduced a reference image, depicting the stationary elements in the scene. Similarly, Karmann and von Brandt [34] used a simplified Kalman filter to predict and update their adaptive background memory. Usual techniques can be applied to detect changes between such a reference image and the current image. In that case, change regions coincide with moving object masks. Such an approach is promising if the background reference is perfect. This method however fails when frequent illumination changes occur, since then updating the reference image becomes a difficult task.

When dealing with small objects in a clutter background, a longer temporal integration is required. Cowart, Snyder and Ruedger [19] used the Hough transform to detect the moving object tracks in accumulated frame difference images. These tracks were assumed to be straight lines. Nakanishii and Ishii [42] use horizontal temporal slices to detect motion and to extract moving vehicles. Given a static camera and assuming constant-speed translational motion, moving areas are extracted from the slice using Hough transform. In some ways, they can cope with crossings or occlusions.

Blostein and Huang [11] have developed a multistage hypothesis testing framework in the spatio-temporal space to detect linear tracks. In practice, the search space has to be reduced to a discrete test set of trajectories (typically ten-frame length), with object velocities within the $[0, 1]$ pixel per frame range. This method assumes a local constant velocity. The motion detection process can be successfully facilitated with the removal of the background structure [11]. Blostein and Huang showed that, if decision is performed considering a long time interval (17 frames in their exemple), randomly fluctuating motions (e.g., trees) are discriminated from the steadily approaching objects.

When the camera is mobile, many approaches imply to compute motion fields, which can become very complex due to motion and depth discontinuities, unknown camera motion, occlusions... Irani et al., [33], have dealt with situations involving camera motions. Moving objects are detected by computing the dominant motion (assumed to correspond to camera motion), and by performing motion compensation using this dominant motion. Tracking of moving objects in the image sequence is based on temporal integration and registration. This approach requires that the size of moving objects is small. Note that compensation techniques have to rely on several hypotheses to make the approach efficient [7], such as validity of affine motion model, planar environment... Other techniques in dynamic scene analysis are more concerned with higher-level approaches, in particular motion-based segmentation schemes based on 2D parametric models [12, 48], or even measurement of the 3D kinematic components of the scene [2, 47].

3 Temporal multiscale decomposition

An important principle of our approach is to derive a temporal decomposition, which allows us to reflect the dynamical content of the sequence at any pixel and at any time. The basic idea is to consider the image sequence as a set of monodimensional time varying signals. A sequence of K images of size $N \times M$ pixels is then a set of $N \times M$ monodimensional signals with K samples. We can quite easily see that previous hypothesis – static acquisition system – is essential, although we desire to tolerate perturbations like oscillations and small motion from the camera. Ideally, in the absence of any illumination variation and any moving object, all these temporal signals can be considered as constant signals: their frequency content is null in that case.

The problem requires an efficient time-frequency transform to characterize the different dynamical components. When detecting moving objects, motion direction and magnitude are not needed. We can thus limit this transform to the temporal axis. Filters of interest here are monodimensional and temporal, and the $N \times M$ temporal decompositions can be carried out in parallel.

3.1 Wavelet transform

Wavelet bases are adopted to ensure a good temporal and frequential localization, [18, 21]. In addition, a broad range of functions are available, allowing us to select the regularity of the basis, to use compact supports, and to implement an orthogonal decomposition. Windowed Fourier transforms, unlike wavelet bases, rely on a constant support envelope, leading to a conflict between temporal and frequential resolutions.

A wavelet basis is composed of a family of functions adjusted by two parameters: one for the position (in time), b, the other for the scale, a. The wavelet basis $\phi_{mn}(t)$ can be written as follows:

$$\phi_{mn}(t) \;=\; a^{-m/2}\, h\left(a^{-m}t - nb\right) \tag{1}$$

and the "time-frequency" representation of $f(t)$ is therefore:

$$\Phi_f(m,n) \;=\; \langle \phi_{mn}, f \rangle \tag{2}$$

Spatial resolution remains here constant over the temporal decomposition levels, whereas in other issues such as object tracking or edge detection for instance [14, 39], multiresolution analysis is applied spatially. The point of view here is quite different.

3.2 Temporal multiscale analysis

Wavelet transform is directly related to multiscale analysis, leading to fast algorithms. We will use the word "multiscale" interchangeably with "multiresolution". The original temporal signal is denoted C^0. We get the following tree structure. At a given level k the signal C^k, called approximation signal, is split up into two terms:

- a new approximation signal at a coarser scale C^{k+1} (H is equivalent to a low pass filter),

$$C^{k+1}(i) = \sum_n H(n-2i)\, C^k(n) \tag{3}$$

- and a signal coding the difference in information D^{K+1} (G is equivalent to a band pass filter)

$$D^{k+1}(i) = \sum_n G(n-2i)\, C^k(n) \tag{4}$$

The sampling frequency decreases as the level becomes coarser (decimation process). During the decomposition process, first computed difference in information D^1 characterizes high temporal frequencies components. Following levels D^2, D^3 ... concern lower frequency bands.

We used several orthogonal bases with compact support, proposed by Daubechies in [21]. The simplest basis is a well known particular case, the Haar basis. Filters H and G support in the latter case are limited to two coefficients:

$$\left\{ \begin{array}{lcl} H(0) &=& 2^{-1/2} \\ H(1) &=& 2^{-1/2} \end{array} \right. \quad \text{and} \quad \left\{ \begin{array}{lcl} G(0) &=& 2^{-1/2} \\ G(1) &=& -2^{-1/2} \end{array} \right. \tag{5}$$

Decomposition formulas are also very simple (image differences and means)

$$C^{k+1}(i) = \frac{1}{\sqrt{2}} \left[C^k(2i) + C^k(2i+1) \right] \quad \text{and} \quad D^{k+1}(i) = \frac{1}{\sqrt{2}} \left[C^k(2i) - C^k(2i+1) \right] \;. \tag{6}$$

Analytic computation of H and G filters coefficients becomes heavy as filters size increases. The general equations for the multiscale decomposition are given in relations 3 and 4. Note that the larger the filter, of course the higher the number of required samples. Therefore, for a filter of size $2 \times Z$, level k can be reached after $2 \times \left[\left(2^k - 1\right) \times Z - \left(2^{k-1} - 1\right) \right]$ input images in our motion detection problem.

For instance, if the chosen support size is equal to six, we need 156 images of the original sequence to get the first level 5 sample of the difference in information signal D^5. Hence, in the experiments reported in this paper, we have limited the multiscale temporal decomposition to 4 or 5 levels, according to the filter size.

Figure 6 shows the temporal multiscale decomposition using Haar basis of the intensity function for a pixel belonging to the horizon line of the so-called "horizon" sequence; it corresponds to a high intensity gradient area. The frame on the left represents the original intensity signal profile over time recorded at that point of the image grid. Then, from left to right we have successively decomposition levels 1 to 5, where top row contains approximation signals C^k, and bottom row difference in information signals D^k. Intensity is coded from 0 to 255 on the vertical axis. We can notice the influence of camera oscillations around its position: these intensity profiles convey temporal fluctuations. We can also note the temporal drift of the camera, slowly tracking the moving plane. As a result, there is an important modification of the considered pixel intensity in Figure 6. For the same sequence, Haar basis representation of the intensity over time of a pixel, located on the projected plane trajectory, is given in Figure 7. The two minima correspond to jet engines passing.

Figure 8 depicts spatial maps of all the D^k signals, obtained at a given time instant for the so-called "horizon" sequence. We successively have levels 1, 2 and 3 on the first row, then levels 4 and 5 on the second row. We can actually see that spatial regularity increases as temporal resolution becomes coarser. Temporal multiscale decomposition thus allows us to avoid a preliminary spatial smoothing, which should be fatal to small size

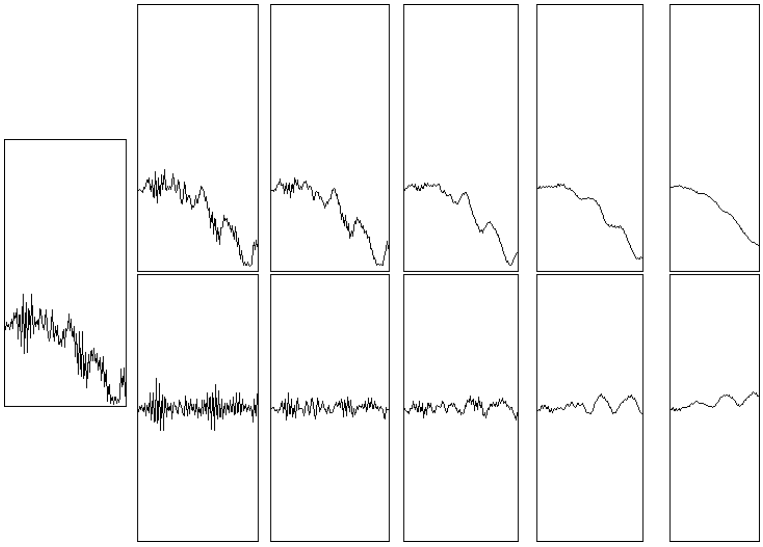

Figure 6: Temporal multiscale decomposition using Haar basis for the "Horizon" sequence, a pixel in the boat area. Top row: approximation signals. Bottom row: difference in information signals. From left to right: original signal then levels 1 to 5.

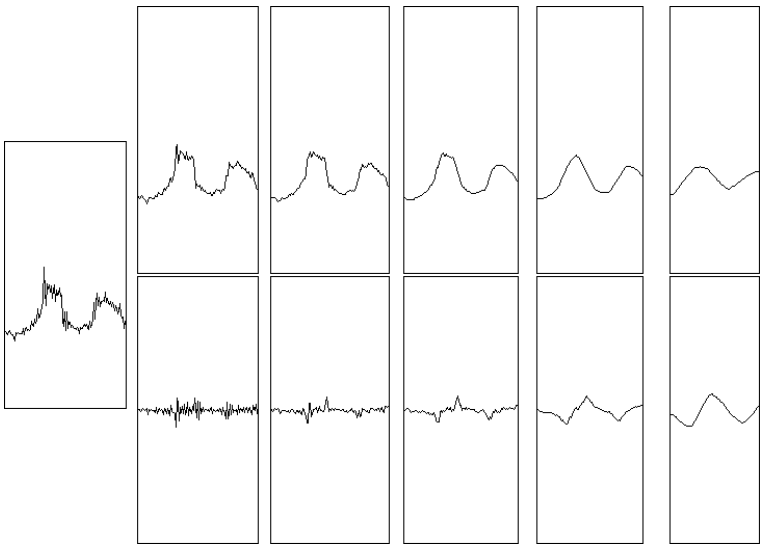

Figure 7: Temporal multiscale decomposition using Haar basis for the "Horizon" sequence, a pixel on the plane trajectory. Top row: approximation signals. Bottom row: difference in information signals. From left to right: original signal then levels 1 to 5.

moving objects in the image plane. These maps are quite noisy, and the plane is hardly visible before the third level. Horizon line can also be clearly seen as time resolution decreases, because of the camera drift.

As a comparison with Figure 7 which used the Haar basis, we have computed the multiscale decomposition for the same temporal signal with a larger filter size. Figure 9 shows a decomposition where H and G have six coefficients. Temporal regularity slightly obviously increases as the filter size is larger. However, we notice that both representations remain similar. Additional computation cost makes the use of a large filter support not legitimate, even more since motion detection requires a fast decision.

3.3 Intermediate decision maps

This multiscale decomposition allows us to build temporal intensity change maps at various temporal scales, i.e., maps resulting from a binary decision such as: `no temporal change` vs. `temporal change` at each temporal

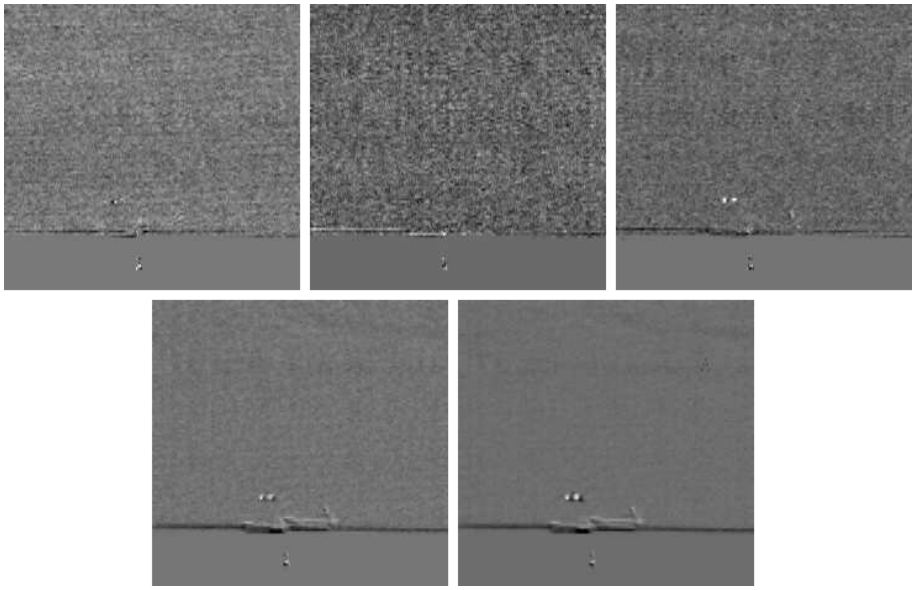

Figure 8: Difference in information maps using Haar basis for the "horizon" sequence. Top row: levels 1, 2 and 3. Bottom row: levels 4 and 5.

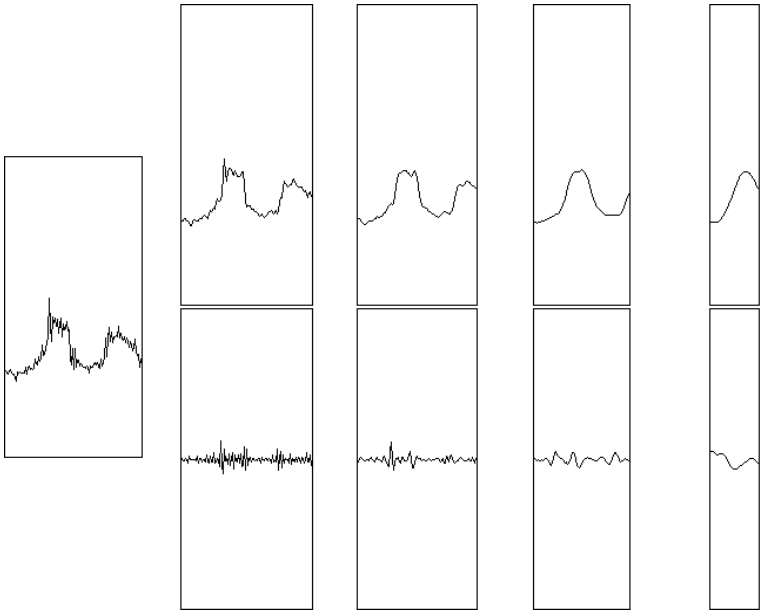

Figure 9: Analogous to Figure 7, but with a filter of length 6 instead of Haar basis.

scale. We propose to give a decision on the presence or absence of temporal changes at each resolution level; we call these binary maps, "intermediate decision maps". Wavelet decomposition gives approximation and difference signals. In the Haar basis case, D^1 signal directly represents a difference between two successive samples of C^0, the original temporal signal. In this way, standard change detection schemes between two images can be applied to binarize the multiscale D^k maps.

3.3.1 Likelihood ratio tests

We apply a two-hypotheses likelihood ratio test to validate or not temporal changes at each scale. This likelihood ratio test is defined as follows, [31]. Two small windows W_1 and W_2 of same size, centred at the same location s but at two successive times, resp. t_1 and t_2, are considered. Within each window, the intensity function is modelled as a deterministic bilinear polynomial function corrupted by an additive zero-mean Gaussian noise

which is supposed to be spatially and temporally uncorrelated. Let x_i and y_i be the relative coordinates of point p_i with respect to the centre s of the window, and let $I_1(p_i)$ be the intensity value at point p_i in image at time t_1, we get:

$$I_1(p_i) = \beta_0 + \beta_1.x_i + \beta_2.y_i + \nu \tag{7}$$

where ν is a zero-mean Gaussian noise of variance σ^2. Let us denote $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2)$. Two competing hypotheses are compared:

- hypothesis H_0 (i.e., `no temporal change` at s):
  the same parameterization θ_0 of the intensity function is valid for both W_1 and W_2.
- hypothesis H_1 (i.e., `temporal change` at s):
  There are two different parameterizations of the intensity function θ_1 in W_1, and θ_2 in W_2.

We use in fact a slightly modified version, introduced in [13], of the original one proposed in [31]. Noise variances are considered as independent of the considered hypotheses, i.e., $\sigma_0 = \sigma_1 = \sigma_2$, and a value σ^2 is computed over each frame as the empirical variance of the frame difference variable. This leads to a simplified version of the test. The log-ratio of the likelihood functions $\mathcal{L}_1$ and $\mathcal{L}_0$ (resp., corresponding to hypotheses H_1 and H_0) is then derived. The decision step can be formalized by:

$$\psi^k(p) \underset{H_1}{\overset{H_0}{\lessgtr}} \lambda \tag{8}$$

where $\psi^k(s)$ is the resulting expression of the log-likelihood ratio after the estimation of parameters θ_0, θ_1 and θ_2 in the maximum likelihood sense. λ is a threshold which may be inferred from tables of statistical laws. H_1 is selected if $\psi^k(p) \geq \lambda$, otherwise H_0 is taken. This is achieved at each considered scale k, and we get an expression of $\psi^k(p)$ in terms of the D^k multiscale maps (N is the number of pixels in the local window):

$$\psi^k(p) = \frac{1}{2\sigma_k^2}\left[\frac{1}{N}\left(\sum_{i=1}^{N} D^k(p_i)\right)^2 + \frac{1}{\sum x_i^2}\left(\sum_{i=1}^{N} x_i D^k(p_i)\right)^2 + \frac{1}{\sum y_i^2}\left(\sum_{i=1}^{N} y_i D^k(p_i)\right)^2\right] . \tag{9}$$

$\psi^k(s)$ follows a χ^2 distribution law with three degrees of freedom. When a scene does not abruptly change, we do not need to compute variance σ_k^2 at each frame. In fact, this empirical value can be considered as constant for a given resolution level k. Therefore, a multiscale decomposition over K levels requires to compute K empirical variance values σ_k^2.

This sub-optimal version [13], although more robust to noise than the original one [31], is still sensitive to illumination variations. However, an illumination variation can often be locally approximated by an illumination offset, i.e., additive modification. Note that Skifstad [45] used this concept to derive a heuristical test based on the divergence of the quadratic model proposed in [31]. In order to accept perturbations of the intensity distribution that can be modelled as an additive constant, we propose to modify the previous linear model. Let us consider a local window whose intensity distribution is zero-mean; this zero-mean intensity function $\bar{I_1}$ (resp. $\bar{I_2}$) can be expressed in terms of the original one I_1 (resp. I_2) as follows:

$$\bar{I_1}(p_i) = I_1(p_i) - \frac{1}{N}\sum_{j=1}^{N} Ip_j \tag{10}$$

The bilinear polynomial which locally models the intensity function is now defined by two parameters: β_1 and β_2. After a few developments, we get a new likelihood ratio test expression, and the decision step can be formalized by:

$$\psi^k(p) = \frac{1}{2\sigma_k^2}\left[\frac{1}{\sum x_i^2}\left(\sum_{i=1}^{N} x_i D^k(p_i)\right)^2 + \frac{1}{\sum y_i^2}\left(\sum_{i=1}^{N} y_i D^k(p_i)\right)^2\right] \overset{H_1}{\geq} \lambda \tag{11}$$

where λ is inferred from the χ^2 distribution law with two degrees of freedom. Hypothesis H_1 is selected if $\psi^k(p)$ is greater than threshold λ

3.3.2 Examples

In real sequences, the dynamical content of the images can become difficult to handle due to sensor noise, misleading changes due to illumination variations, occlusions... As an illustration, we give intermediate decision maps corresponding to the two original sequences, presented in Section 1. Note that the threshold λ in the likelihood ratio test is kept the same for all sequences we have tested, i.e., $\lambda_{\chi^2_{.01,2}}$ (corresponding to a 99% confidence). In any case, intermediate decision maps barely change when confidence threshold is slightly modified.

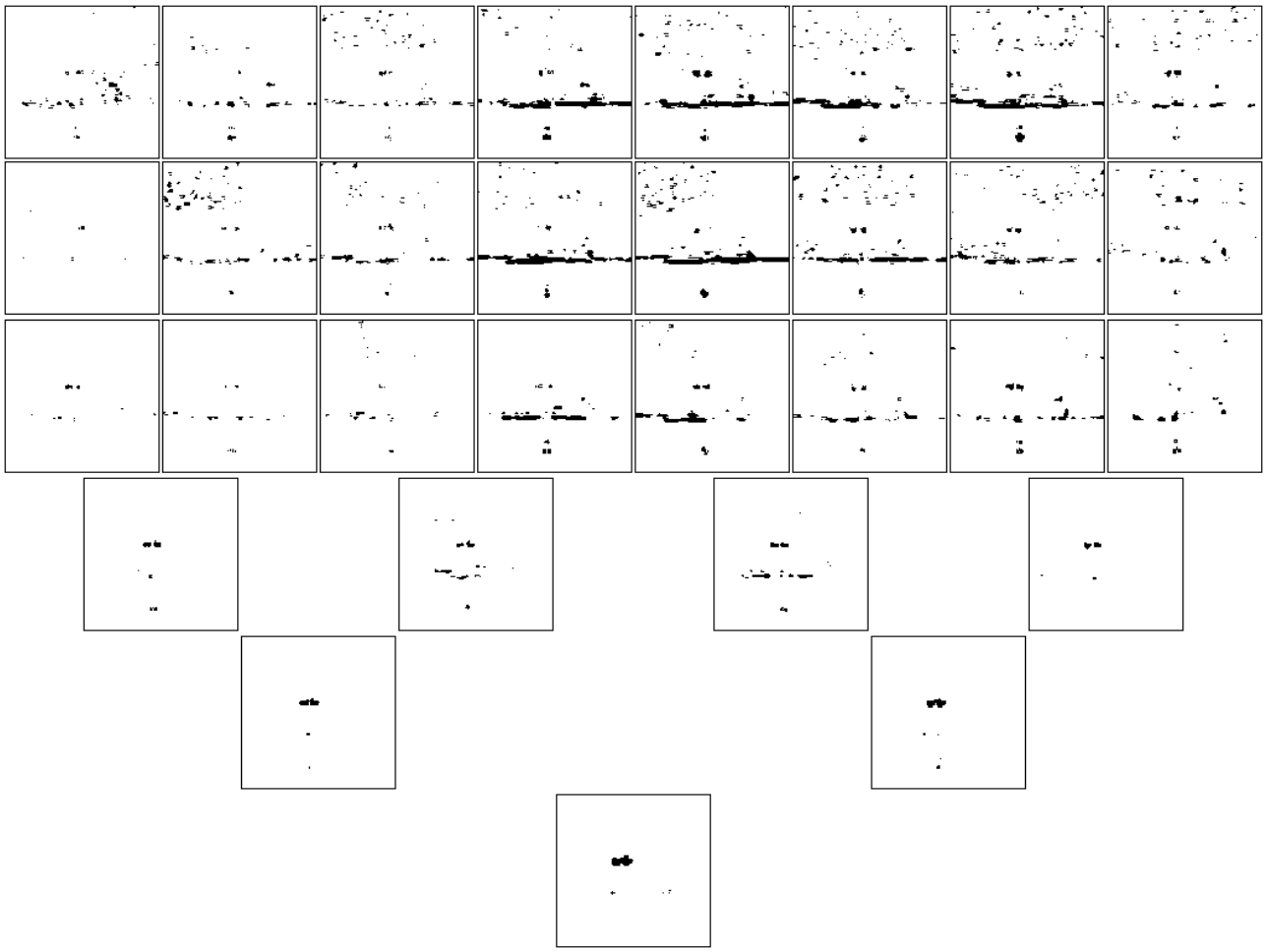

Figure 10: Pyramid of intermediate decision maps using the Haar basis of "horizon" sequence. Layout explanation is given in Figure 11.

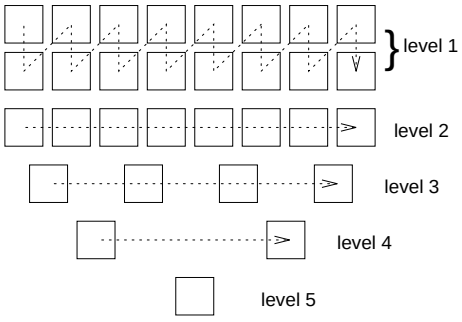

Figure 11: Layout of the pyramid of maps of Figure 10.

Figure 10 shows a complete temporal decomposition pyramid on five levels of the "horizon" sequence. This process requires 32 input images as we use the Haar basis. The pyramid is computed every second image, so

that the decision process rate is half the original sequence temporal sampling rate; the temporal multiscale decomposition is carried out on a sliding temporal window. Camera oscillations are visible at levels 1 and 2 – some areas around the horizon line are detected as `temporal change` – because of the high frequential components of this kind of perturbation. However, the same points located near the horizon line that reacts at coarser levels are more likely due to the low tracking movement of the camera. We notice that the moving plane mask stretches along its trajectory as temporal resolution decreases. At level 5, both jet engines even form a connected blob, because of the horizontal direction of the apparent plane motion. First level (two top rows of Figure 10) represents nothing but the likelihood ratio test performed between two successive images, [13].

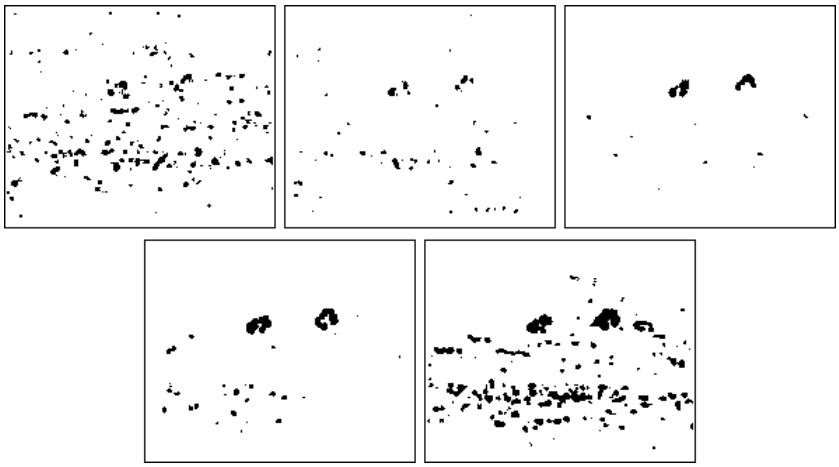

Figure 12: Intermediate decision maps using the Haar basis corresponding to the "bush" sequence at $t = 32$. From left to right, top row: levels 1, 2 and 3. Bottom row: levels 4 and 5.

Intermediate decision maps at time $t = 32$ of the "bush" sequence are shown in Figure 12. In these maps, like in Figure 10, the locations detected as `temporal change` that correspond to perturbations, i.e., that could lead to false motion detection, are not correlated between levels. In other words, the misleading `temporal change` detections caused by perturbations, do not occur at the same location on successive levels. On the other hand, once a moving object is detected at a given level, it is also detected as `temporal change` at every coarser level, and the detected mask lengthens along the object trail. If we only used one temporal resolution level, i.e., the original image sequence sampling rate, we would not be able to discriminate "true" motion from spurious detections due to various perturbations (e.g., illumination variation, parasitical motions): indeed, temporal change maps at level 1 are very corrupted.

Figure 13 contains the intermediate decision maps for an infrared sequence of a countryside landscape, corrupted by something equivalent to an illumination variation. This scene will be referred as the "illumination" sequence. There is a road in the foreground going away to the upper-right corner of the image (see original images in Figure 25). Two trucks far away in the background move from right to left (see motion detection masks in Figure 25). In order to get more contrasted data around time $t = 50$, camera aperture is increased progressively over an interval of 30 images, changing illumination effects. The vehicle on the left cannot be seen before the modification of the camera aperture. Figure 13 demonstrates the benefit of considering a zero-mean intensity function to model the local intensity distribution in the temporal change detection tests (see Subsection 3.3.1). When illumination can be considered as stable, the change detection test based on a complete linear model, as developed in [13], is reliable. Indeed, the two moving vehicles in the "illumination" sequence are nicely recovered from level 2 to level 5: this is shown by the maps on the left-hand side of Figure 13. However, during the modification of the camera aperture, the required assumption of constant illumination is no longer valid, and the test based on a complete linear model significantly deteriorates, as it can be seen in the central maps of Figure 13. On the contrary, the temporal change detection approach that we have proposed in Subsection 3.3.1 is capable to successfully cope with illumination variations that can be locally considered as additive perturbations. The improvement is quite substantial, and the moving vehicles of the "illumination" sequence can be recovered even during the illumination variation time interval: see level 2 to 4 of the maps on the right-hand side of Figure 13. Let us note that we still have some irrelevant temporal change detections at level 5 around areas mostly corresponding to strong spatial gradients of the intensity.

These experiments show that we can take advantage of a temporal multiresolution approach. Now, we have to improve the detection stage by combining in a proper way these various types of temporal change information.

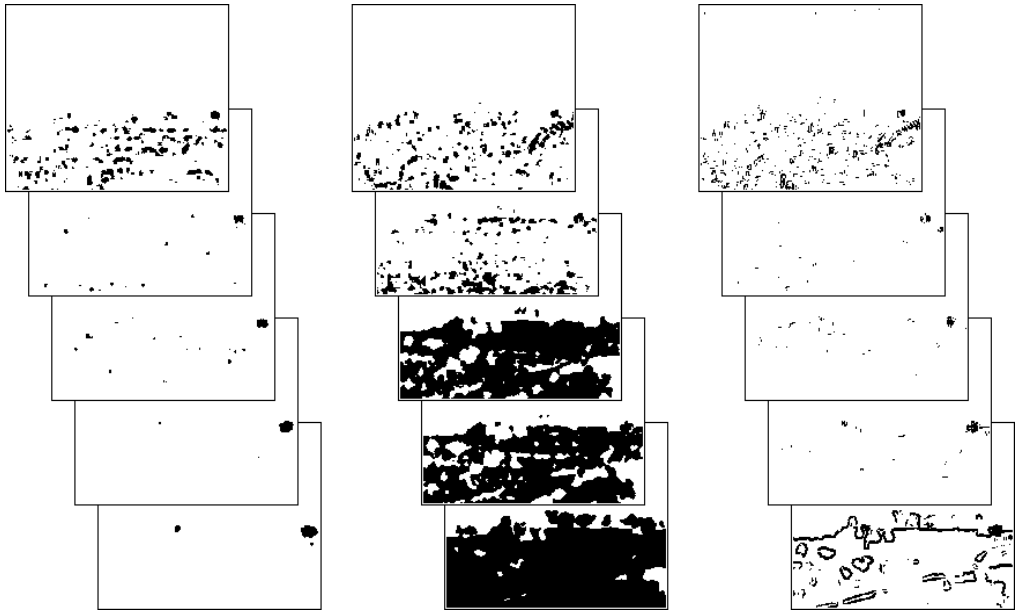

Figure 13: Intermediate decision maps using the Haar basis corresponding to the "illumination" sequence. Comparative experiments concerning the temporal change detection tests: the first one is based on a complete linear model (see [13]), denoted A, and our test based on a zero-mean intensity distribution model, denoted B. From top to bottom : levels 1 to 5. Maps on the left: test A before the modification of the camera aperture. Maps in the middle: test A during the illumination modification. Maps on the right: test B during the illumination modification.

3.3.3 Qualitative characterization of typical behaviours

Information about the dynamical components we encounter in the intermediate decision maps can be summarized as follows:

- most environmental static areas are detected as `no temporal change` at the various resolution levels;
- sensor noise leads to isolated `temporal change` detections;
- as soon as a moving object is detected at a given temporal scale, this object is still detected at any coarser temporal scale;
- perturbations (like those due to small camera oscillations) generally cause `temporal change` detections over a limited number of levels.

The decomposition process has to be limited in practice to a given number of temporal levels. We propose to set the temporal decomposition to five scales. This has proven to be a good trade-off between algorithm efficiency and computation load. We have observed that a decomposition on at least five levels is required to allow us to correctly discriminate the various dynamical behaviours present in the scene. To illustrate the three specific dynamical behaviours we have characterized, we have derived the following heuristic rules for a multiresolution decomposition on five levels:

- if a pixel at time t is at least detected as `temporal change` at three successive temporal scales, it has a "high probability" to belong to a moving object;
- if a pixel at time t is never detected as `temporal change` at any temporal scale, it can be considered as static;
- if a pixel at time t is at most detected as `temporal change` at two successive temporal scales, it is likely to belong to areas corresponding to "parasitical" temporal change.

Figure 14 illustrates the previous set of rules applied to the temporal decomposition of a given temporal signal. The top part represents intermediate decision vectors coresponding to the following binary decision