EURASIP Journal on Applied Signal Processing 2002:6, 582–594

Tracking of Objects in Video Scenes with Time Varying Content

Amal Mahboubi
IRCCyN UMR 6597 CNRS EPUN, rue Christian Pauc La chantrerie, BP 60601, 44306 Nantes, France
Email: amal.mahboubi@polytech.univ-nantes.fr

Jenny Benois-Pineau
LABRI, CNRS UMR 5800, Université Bordeaux-1, 33405 Talence, France
Email: jenny.benois@labri.fr

Dominique Barba
IRCCyN UMR 6597 CNRS EPUN, rue Christian Pauc La chantrerie, BP 60601, 44306 Nantes, France
Email: dominique.barba@polytech.univ-nantes.fr

Received 27 July 2001 and in revised form 5 February 2002

We propose a method for tracking of objects contained in video sequences. Each video object is represented by a set of polygonal regions. A bottom up approach (spatial segmentation/motion estimation) is applied for the initialisation of the method, a limited human interaction is used to build the semantic map of the first frame in video sequence. The tracking of this model along a video sequence is based on detecting and indexing new objects in a video scene. Semantic rules are used to label new objects and, the current state of segmentation is validated by forward projection of the background.

Keywords and phrases: new region extraction, labelling, object indexing, forward projection.

1. INTRODUCTION

The new ongoing standard of video representation and coding MPEG4 [1] gives tremendous possibilities for the composition of heterogeneous video scenes combining video objects of various nature. The main challenge behind MPEG4 technology is the development of efficient and truly automatic methods for extracting and tracking of objects in video. Once video objects are known at each time instant, they can be manipulated, put into another scene and so on. Numerous research works, developed recently [2, 3, 4, 5, 6], are devoted to the problem of automatic tracking of a selected video object planes (VOP) in a scene. The focus of study [2] is an object tracking which combines motion and spatial information in order to be able to track objects which do not present either homogeneous texture or motion. The method has been applied to the problem of the generation of video objects for content-based functionalities in object-based coding schemes. To extend this work, [3] presents an interesting technique for generic object tracking. This method perfects the first tracking scheme [2] by projecting region-based partition. This projection accommodates the previous partition information in the current image. Then the object partition is re-segmented and it is projected on the following image using motion information. Nevertheless, this method assumes that objects have been defined in the first image and the process is not able to correctly detect a new object in the scene. To mitigate those consequences, it introduces the concept of user interaction in the algorithm to handle variation of objects. The works [4, 5] surveyed the region-based active contours approaches. The first study introduces the active contour criteria and the second presents the use of a B-spline parametric contours for object tracking. The temporal gradient is a term of temporal evolution of scene content. Here the principle is based on applying the force based over few contour points, then contour evolution depends only on the evolution of these interpolated points and no more on each point of contour.

The study [6] suggested a backward tracking technique. The principle of this method is the use of a spatial segmentation on each frame. This segmentation is then back-projected according to the motion of each region onto a reference frame where the initial segmentation is labelled as *object regions* and background. It results in a good localization of the boundaries as they are obtained by an intra-frame segmentation. The method also shows a good capacity to follow the deformation of objects. The spatial segmentation used is based on the minimum description length criterion. Motion compensation is performed using affine motion model per region estimated with a multiresolution scheme and relaxation. The encountered problems of this method are the possible artefacts in spatial segmentation and errors in motion estimation and in the occlusion areas as well. To remedy to that, a splitting of regions is proposed to correct these artefacts according to the object definition or prediction error.

But other faces of spatio-temporal segmentation remain unstudied. The main one is an automatic semantic interaction between different areas in a natural video scene, that is, the ability to correctly label as *object, new object, background* all areas in video scenes with strong changes between successive time instants. In this context, we propose a tracking of objects in video scenes with time varying content.

The method starts with the extraction of video object from a complex natural video scene at the initial time instant, using a fine spatial partition of image plane. The geometry of each spatial primitive is represented by a piece-wise linear approximation of the border. Affine motion model of each polygonal region is estimated by means of gradient descent method. Then all regions are classified semantically by means of human interaction. After that, connected regions are merged in each semantic class to build a hierarchical representation of the scene using motion homogeneity criteria. The tracking of changed content is based on motion estimation of regions along the time and on textural and topological coherence measures.

Correction or confirmation of labelling for the current frame is based on a forward projection of the background taken at specified moments in the sequence.

The paper is organised as follows. Section 2 describes the initialisation of object-based partition of video scenes. General tracking scheme is described in Section 3. Section 4 represents the indexing of VOPs in case of time-varying content. Section 5 describes the validation of the automatic labelling. Finally, the main results of the tracking are presented in Section 6.

2. INITIALISATION OF SPATIO-TEMPORAL PARTITION OF VIDEO SCENES

To extract objects to be tracked, a spatial colour-based segmentation of a video frame at the initial time instant is applied. The spatial segmentation is the result of a morphological method based on modified watershed [7]. It consists of four classical steps of a morphological scheme: image simplification, gradient computation, marker extraction, watershed algorithm. In our approach we especially developed aspects of the watershed method. After image simplification using reconstruction by opening-closing filters, morphological gradient is computed [7, 8]. The result of this step is the gradient image which highlights the grey level contours contained in the filtered image. Then the marker is extracted, the result here is a binary image (marker image) where 1 represents the pixels with a gradient magnitude lower than a fixed threshold. Finally, region growing by watershed in colour YUV space is fulfilled. Namely, each connected area in the marker image is labelled. It represents a seed of future region characterised by its mean colour vector $(\bar{y}, \bar{u}, \bar{v})_i^T$. Then the seeds are expanding absorbing pixels on their border if their colour distance from mean vector $(\bar{y}, \bar{u}, \bar{v})_i^T$ is less than prefixed threshold. This threshold depends on mean colour values of region. After all pixels have merged for a given threshold value for each region, the relaxation process with corresponding thresholds allows for further merging until all pixels are labelled in image plane.

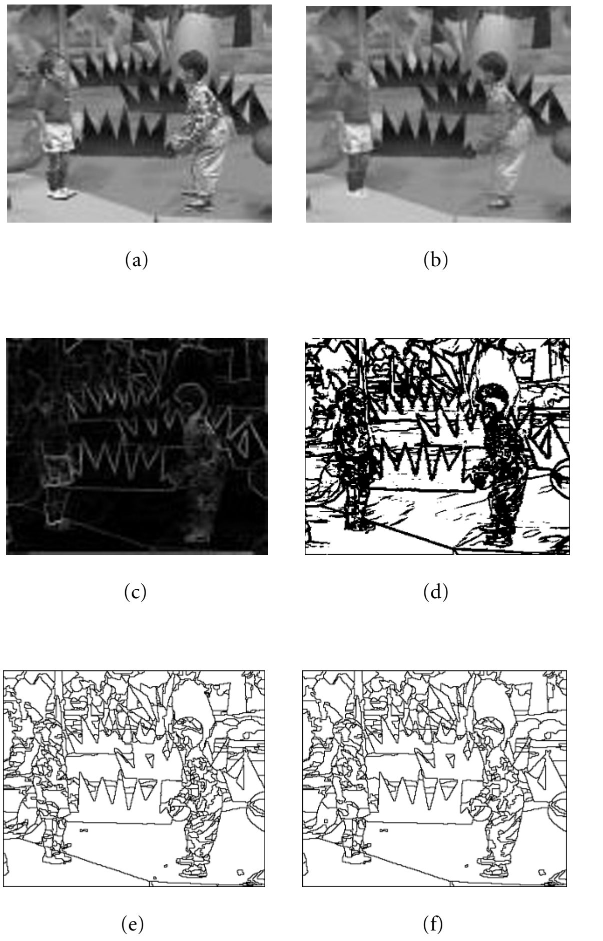
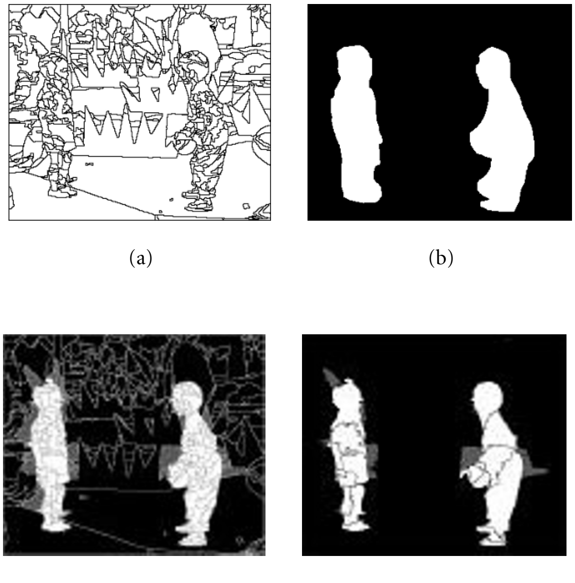


FIGURE 1: The process of the spatial segmentation. (a) Original frame, (b) filtered image, (c) gradient image, (d) marker image, (e) watershed image, (f) final image.

The result of the watershed segmentation is too redundant as it gives a very fine partition containing several regions. Therefore we apply a hierarchical merging based on a relative contrast criteria. Adjacent regions of the final result represent a rich partition of image plane [8]. Figure 1 displays the different steps of the spatial segmentation.

For each spatial region a polygonal representation is constructed using a piece-wise linear approximation of its border. To build the spatio-temporal structure, we estimate the motion of each polygonal region by the gradient descent method [9]. Here a global affine motion model is supposed



(a) (b)

(c) (d)

Figure 2: The process of the first semantic classification. Sequence "Children" frame at $t = 3$ (a) spatial segmentation, (b) user mask, (c) first result of the semantic classification, (d) final result of the manual semantic classification and motion-based merging (black for Background, white for the VOPs and grey for Uncertain).

with parameter vector $\Theta = (t_x, t_y, k, \theta)^T$. According to it, an elementary displacement vector $(dx, dy)^T$ at each pixel position (x, y) in a given region is expressed as

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} tx \\ ty \end{bmatrix} + \begin{bmatrix} \text{div} & -\text{rot} \\ \text{rot} & \text{div} \end{bmatrix} \begin{bmatrix} x - x_g \\ y - y_g \end{bmatrix}. \tag{1}$$

Here x_g, y_g are the coordinates of the gravity centre of the region, tx, ty, are translation parameters, *div* is a zoom parameter, and *rot* is a rotation parameter [10].

These regions should be labelled semantically to provide VOPs corresponding to meaningful objects in a scene. Purely automatic labelling is possible only for simple scenes, where a strong difference of the dynamic range and of the textural characteristics of objects and the background is observed. In general case of natural scenes, an object can be partly static and thus, it cannot be distinguished from the background based on motion difference. The colour and texture of object can be similar to the background. Therefore, a user interaction is required to completely extract objects in a general case. We propose a minimal human intervention. The user creates a binary semantic class mask on the first frame by encircling objects, here we have an image with 0 in the background and 1 inside objects (the encircled area) as shown in Figure 2b. This binary image called *user mask* is then used for the initial VOP labelling.

Each polygonal region (Figure 2a) is superimposed on the user mask (Figure 2b) to get the initial classification (see Figure 2c). Thus the three semantic classes can be introduced:

(1) *Object*: is the class of objects in the scene.
(2) *Background*: this class denotes generally the scene background.
(3) *Uncertain*: this class represents the ambiguous area on VOPs borders.

The semantic labelling of each region is based on the ratio of region pixels bellowing to object area in user mask. Denote by ΩR_i the set of region pixels which are labelled as object in user mask: $\Omega R_i = \{p_i \in R_i / \text{Mask}[p_i] = 1\}$. Also denote by ΓR_i the set of region pixels belonging to the background $\Gamma R_i = \{p_i \in R_i / \text{Mask}[p_i] = 0\}$. Then, the semantic label of R_i is assigned as follows:

$$\begin{aligned} \text{Form}(R_i) &= \begin{cases} 1 & \text{if } \left(\dfrac{\text{Card}(\Omega R_i)}{\text{Card}(R_i)}\right) > \text{Th}_{\text{obj}}, \\ 0 & \text{otherwise,} \end{cases} \\ \text{Back}(R_i) &= \begin{cases} 1 & \text{if } \left(\dfrac{\text{Card}(\Gamma R_i)}{\text{Card}(R_i)}\right) < 1 - \text{Th}_{\text{obj}}, \\ 0 & \text{otherwise,} \end{cases} \\ \text{Unce}(R_i) &= \begin{cases} 1 & \text{if } 1 - \text{Th}_{\text{obj}} < \dfrac{\text{Card}(\Omega R_i)}{\text{Card}(R_i)} < \text{Th}_{\text{obj}}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

Here Th_{obj} is the ratio of region pixels inside object mask with regard to the whole number of region pixels; Card denotes the cardinal of pixel set. An example of this semantic labelling is given in Figure 2c.

The labelled spatial regions constitute a fine partition of the image plane which is too redundant with regard to the scene content (see Figure 2a). Therefore, a motion-based merging process is necessary to construct more meaningful region-based partition. We follow the motion-based merging strategy proposed in [11] to construct a nested hierarchical polygonal partition inside each semantic class (see Figure 2d).

Finally, each VOP is indexed in the video scene by the following method.

Each polygonal region in the image plane corresponds to a region-node in the region adjacency graph (RAG). Starting from an arbitrary Object Class node in RAG, all the graph is traversed by the so-called *In-Depth Search* algorithm [12] and the maximal subgraph with only Object Class nodes is isolated. This subgraph corresponds to a connected VOP in the scene. All region-nodes of this subgraph receive a label, *Object Index*. The process is reiterated for all remaining Object Class nodes with incremented Object Index label.

Resulting from this process, the label of each region in image plane partition is set to Uncertain, Background or its own Object Index value corresponding to the VOP index.

3. TRACKING SCHEME

The principle that guides our tracking scheme was developed in [10, 13, 14] for polygonal partition of video frames. Based on affine motion model of 2D apparent motion (1), it consists in the following steps:

(i) projecting of polygons in the direction of time axis,
(ii) adjusting of predicted borders by an active contour model,
(iii) segmenting of regions with changed content, and
(iv) merging of regions at time $t + 1$.

Thus the spatio-temporal partition S^{t+1} is obtained from S^t. In scenes with changing content, it is necessary to label new regions as belonging to new or preexisting VOPs, to the Background or to Uncertain class. The method presented here incorporates solutions for labelling the new regions and correcting errors of segmentation due to diverging tracking process in the case of strong motion. The tracking of polygonal partition proposed in [10] can yield the appearance of new regions between times t and $t + 1$, which receive new labels. There are two reasons for the creation of new regions in the segmentation map.

When projecting a spatio-temporal partition S^t to the next frame at step (i) with affine model (1), overlapped and uncovered areas are formed in image plane at time $t + 1$. In our previous work [10], we studied in detail processing of occlusions in overlapped areas. For these occlusions, their motion-based assignment to already existing regions was proposed. Thus they do not yield new regions. Another situation is observed in case of uncovered areas, which do not have a prehistory. They can appear in the neighbourhood of VOPs. They also appear on the borders of video frames in case of background motion. They can also be observed inside VOPs (self-uncovered areas).

The second reason of generation of new regions by tracking method [14] is the motion-based segmentation of regions with increased motion compensation error (step (iii) in tracking scheme). In fact if the error of motion compensation with affine model (1) increases between t and $t + 1$, it can be supposed that the given region is not well described by a single motion model. Therefore, it should be split into smaller regions homogeneous with regard to chosen motion-based criterion. Much details relative to this phase are given in Section 4.2. Thus the problem here is to correctly label split regions.

The *uncovered* and *split* regions contain both parts of new objects or preexisting objects and of the background. They are to be labelled with Object Index value, Background, or Uncertain labels. We show in Section 4 how this goal can be achieved.

The final step (iv) of tracking scheme consisting in merging regions is necessary to reduce the redundancy of segmentation, but it can yield segmentation errors in case of weak relative motion of objects and the background. Therefore in this work we add the validation of current segmentation with its state in the past including the initial state, when semantic labelling is based on human interaction.

4. NEW REGION LABELLING

As we noted above, new regions issued from uncovered areas and also from motion-based splitting. In order to correctly label these regions we propose two different approaches. The labelling of an uncovered region is based on the texture analysis of its spatio-temporal neighbourhood and the label distribution of its support in the past. The labelling of a split region is based on motion analysis.

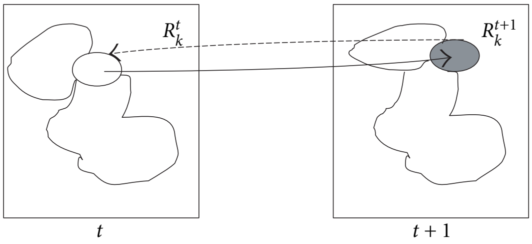


FIGURE 3: Diagram of a back-projection.

4.1. Uncovered regions

Regions in uncovered areas formed by projection of segmentation, can be adjacent to VOPs borders, or to be situated inside an articulated VOP. To label these areas, two measures are combined: a score of pixels belonging to a specific class (Object, Uncertain, Background) in the past reference frame on the one hand, and a texture similarity measure in the current frame on the other hand. These two measures are mixed in one decision rule.

The first measure denoted *Score* refers to the class to which each pixel of region R^{t+1} back-projected into frame I^t does belong. The second measure denoted L indicates the class of a region in the neighbourhood of R^{t+1} in the current frame, which has the most similar texture to the texture of R^{t+1}. Trust weights are assigned to each of these measures and the resulting class label for the region R^{t+1} is that maximising the global trust measure.

Denote by R_k^{t+1} a new uncovered region in frame at t, with its motion parameters θ_k^{t+1} computed in backward direction according to the model (1). Denote by R_k^t the back projection of R_k^{t+1} into image plane at t realised with motion parameters θ_k^{t+1}, as is shown in Figure 3. (Note that the displacements (1) can yield the projection of a pixel of R_k^{t+1} into an inter-pixel position. Then the nearest pixel is taken.)

Denote by $O^T = \{o^t(x, y)\}$ the observation filed corresponding to the value of semantic label in pixel (x, y) in frame at t. Then the score of region R_k^{t+1} with regard to the class C_j, $j = 1, \dots, 3$ (Object, Background, Uncertain) in frame at t will be

$$\text{Score}\left(\frac{C_j}{R_k^{t+1\to t}}\right) = \sum_{(x,y)\in R_k^{t+1\to t}} \delta\big(o(x, y) - \omega_j\big). \tag{3}$$

Here $\Omega = \{\omega_j, j = 1, \dots, 3\}$ if the class label corresponding to Object, Background, Uncertain,

$$\delta(a - b) = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The importance of this score for the given region can be

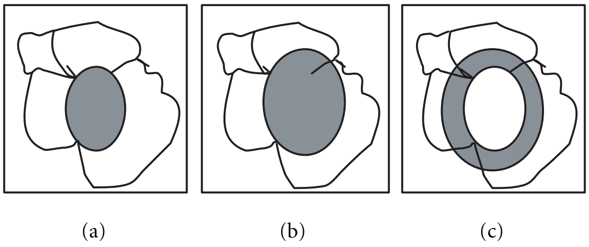


Figure 4: Diagram of uncovered region and its neighbourhood.

measured by its weight:

$$\text{Weight}_{S_j} = \frac{\text{Score}\,\left(C_j/R_k^{t+1\to t}\right)}{\text{Card}\,\left(R_k^{t+1\to t}\right)}. \tag{5}$$

Here $\text{Card}(R_k^{t+1\to t})$ is the cardinal of the set $R_k^{t+1\to t}$.

The computation of texture similarity measure is based on assumption of Gaussian grey-level distributions in limited windows surrounding the given region. These windows are constructed by dilating the region R_k. The assumption here is that the region R_k^{t+1} most likely belongs to such a class (Object, Background, Uncertain) with which it has the most similar texture. Supposing Gaussian distribution of grey-level value inside the region R_k^{t+1}, we will also suppose Gaussian distribution in limited windows surrounding the region. For each region R_{ki}^{t+1} connected with R_k^{t+1} in region adjacency graph, the window W_{ki} is defined as

$$W_{ki} = \delta_\varepsilon \circ R_k^{t+1} \cap R_{ki}^{t+1}. \tag{6}$$

Here δ_ε denotes the morphological dilation operator with structured element of radius ε.

Figure 4 depicts the method: in Figure 4a the region is denoted by hatched pattern, Figure 4b presents a dilated region, the resulting windows are shown in Figure 4c (hatched pattern).

Thus the parameters of windows $\{W_{ki}\}_{i=1}^N$ in neighbourhood of R_k will be mean μ_{ki} and variance σ_{ki}^2. The neighbour likelihood τ_{ki} is computed as (see [13])

$$\tau_{ki} = \frac{\left(I(x,y) - \mu_{ki}\right)^2}{2\sigma_{ki}^2}, \quad (x,y) \in R_k^{t+1}, \tag{7}$$

where $I(x,y)$ is the grey level of the pixel (x,y).

It expresses the hypothesis of the same Gaussian distribution of the grey level both in R_k^{t+1} and W_{ki}^{t+1} in its neighbourhood.

Now we introduce the likelihood of a class $C_j = 1\cdots 3$ (Object, Background, Uncertain) as

$$L_{kj} = \max_{i:\Omega(R_{ik}^{t+1})=j} \tau_{ki}, \quad i = 1,\ldots,h,\ h \le N. \tag{8}$$

For the given class C_j there is no region of this class in the neighbourhood of R_k^{t+1}, then $L_{kj} = 0$.

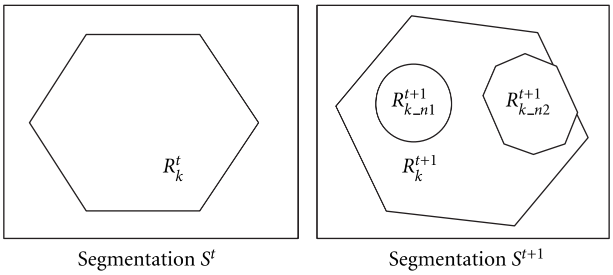


Figure 5: Diagram of split region.

The *amplitude* of this similarity with regard to other classes can be expressed by its weight:

$$\text{Weight}\,L_j\left(R_k\right) = \max\left(\frac{L_{kj}}{\sum_{j=1}^3 L_{kj}}\right). \tag{9}$$

Finally, the class label of the region is assigned according to the maximum of the mixed function of its past labels and texture similarity:

$$\text{Label}\,\left(R_k\right) = \text{ArgMax}\,\left(\alpha\,\text{Weight}_{S_{kj}} + (1-\alpha)\,\text{Weight}_{L_{kj}}\right), \quad j = 1\cdots 3. \tag{10}$$

The coefficient α allows for privileging one measure over another. In the actual study it was set to 0.5.

4.2. Split regions

In order to introduce the method for labelling split regions we will describe how these split regions are obtained in general tracking scheme.

When a region R_k^t is re-segmented at time $t+1$, the resulting set of regions $\{R_k^d\}$ is called split regions (see Figure 5).

The split method (iii) is based on a motion criterion using Markov random field modelling [14]. Firstly, we select regions having a significantly large surface ($\text{Size}(R_k) > \text{Th}_{\text{size}}$) then we study the increase of the mean square prediction error (MSE) inside the regions. To do this a morphological filtering is applied to the region mask before the MSE computation in order to exclude the influence of occluding borders. If this MSE is higher than the splitting threshold ($\text{MSE}(R_k) > \text{Th}_{\text{split}}$) then the motion-based label [15] segmentation method is applied to split the region.

Globally, the MSE increase indicates a content change, unfortunately the significant value of this measure on a given region can be due to the incompatibility of the used model of motion with 2D apparent motion in image plane. To remedy to that we observe the MSE in the original region and its subregions after the segmentation. If the MSE is greater than the sum of the MSE on the n_d subregions then the segmentation can be accepted,

$$\text{MSE}\,\left(R_k\right) > \sum_{i=1}^{n_d} \text{MSE}\,\left(R_{k_i}\right). \tag{11}$$

The motion-based segmentation method consists in the optimisation of the energy functional: $U(O, E, \theta, n) = U_1(E) + U_2(O, E, \theta, n)$, where U_1 corresponds to the a priori modelling of label field and U_2 to conditional likelihood knowing E, θ, and n.

Here $O = \{o(x, y)\}$ is the observation field corresponding to the motion compensation error in each pixel (x, y), $E = (e(x, y))$ is the label field, θ are motion parameters.

The method includes three phases:

(1) detection of a region to split (segmentation of region which satisfies th_{size} and th_{split}) and estimation of their parameters;

(2) initialisation of segmentation map: detection of new regions inside the given one. Here for the introduced motion model, the square error of motion compensation in each pixel is compared to a threshold. The ill-estimated pixels are regrouped into connected components and the motion parameters are re-estimated. The process is reiterated until the stability of labelled regions;

(3) optimisation of the segmentation map (estimation of the optimal label field E). For all the pixel-sites in the regions issued from the first step, the $\tilde{r}$ label supplying the minimal energy in the neighbourhood of each pixel-site s is found:

$$\triangle U_s(\tilde{r}) = A \sum_{(s,l)} \left(1 - \delta(e_l, \tilde{r})\right) + \left[I^{t+1}(s) - I^t\left(s + \vec{d}\left(\theta(\tilde{r})\right)\right)\right]^2. \tag{12}$$

Here l is the label of pixel in the neighbourhood of pixel-site s, $\vec{d}$ the elementary displacement according to (1).

If the best candidate label r supplies high local energy, then the label is re-assigned. The process is reiterated. The optimisation is realised by ICM method.

When the resulting split regions are constructed, the problem is to define which of them corresponds to a new moving object superimposed on the preexisting background or to a new detail in the preexisting object. The method we propose is based on the measurement of a differential motion activity of each split region. The assumption here is that a new significant region belonging to a new object strongly changes its motion between two successive frames. Here, to measure this activity it is necessary to compute motion vector of each pixel of region at time $t + 1$ and of the same pixel of region at time t. Let r_{ki}^{t+1} denote a subregion resulting from motion-based segmentation of region R_k^{t+1} at time $t + 1$. Let θ_k^t be the motion parameter vector of R_k at time t, θ_{ki}^{t+1} is the motion parameter vector of r_{ki}^{t+1}. If the region r_{ki}^{t+1} is back-projected into the image plane at time t, then to the pixel position (x, y) at $t + 1$ corresponds the position $(x + dx, y + dy)$. The elementary displacement vectors $\vec{d}(x, y, \theta_{ki}^{t+1})$, $\vec{d}(x + dx^{t+1}, y + dy^{t+1}, \theta_k^t)$ are computed at time $t+1$ and t for each pixel position (x, y) and $(x+dx, y+dy)$, respectively, with motion parameters of the regions r_{ki}^{t+1} and R_k^t.

Then the measure of differential motion activity we introduce is expressed as

$$\triangle_{mvt}\left(R_{ki}^{t+1}\right) = \frac{1}{\text{Card}\, R_{ki}^{t+1}} \times \sum_{(x,y)\in R_{ki}^{t+1}} \left\| \vec{d}\left(x, y, \theta_{ki}^{t+1}\right) - \vec{d}\left(x + dx_{t+1}, y + dy_{t+1}, \theta_k^t\right) \right\|^2. \tag{13}$$

If this measure is stronger than the activity threshold, then the region R_{ki}^{t+1} is labelled as Object-class region. Such a labelling corresponds to the assumption that objects can strongly change there motion but the background cannot do it.

5. CONFIRMATION OF SEGMENTATION BY THE PAST

Errors in motion estimation and errors in merging regions represent the risk of an automatic tracking scheme. To improve labelling, we introduce the forward bringing process, this study is similar to the recent literature [16, 17, 18, 19] on mosaic image construction from video. A review of literature [16, 17, 18] shows that a mosaic image can be seen as a summary representation of the video. The work [16] defines the mosaic image as the global view of the scene background resulting from camera motion compensation. The study [17] presents a mosaicing methodology which overcomes restrictions on the different types of camera motion (translating sideways, panning camera, or both) by using a manifold, where shape is determined adaptively based on the motion of the camera during the mosaicing process. The survey [18] suggests the minimisation of the alignment error between the already defined mosaic image and the current image to get a better reconstruction.

Generalising the principle of mosaicing, it can be expressed as bringing all pixel values in video sequence into one image plane by motion compensation.

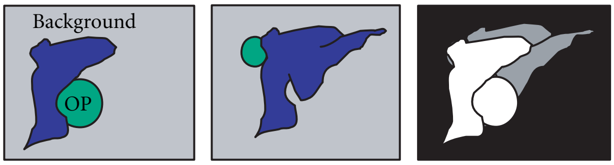
This principle in its simple form (2D affine motion (1)) will be used now to confirm or correct the segmentation at a current moment of time, using its state in the past. At the beginning of a video sequence, the semantic labelling of region (Object, Background, Uncertain) is based on human interaction. Object and Uncertain regions being excluded, the background set of pixels in the first frame is *certain,* in the sense that it corresponds to user interpretation. This certain background will now be used in segmentation correction by the *forward brining* process, which includes two steps:

(1) certain background projection;

(2) segmentation correction.

5.1. The forward projection of the background

Consider the background at the time instant when it is certain noted Back_0 and the segmentation map S^t at time t, $t_0 < t < t_n$. The problem now is to project the background Back_0 at time t and to superimpose it on the map S^t. This forward projection is done recursively using the motion parameters of appropriate Background region according to the



Segmentation t_0 Segmentation t_n Back_0 projected at t_n

FIGURE 6: Forward projection of Back_0 diagram.

tracked label map $S^0, S^1, \dots, S^n$. Each point p of Back_0 is projected with the motion parameters $\overrightarrow{\theta}_R$ of its own region according to the tracked label map when the region label is *Background* (see the black area in the right image of Figure 6). If at an intermediate time instant $t_0 < t < t_n$ the projected pixel p is not labelled as Background in a segmentation map S^t, then we take the parameters of the dominant background region in S^t to project pixel p onto the image plane at time $t + 1$ (see grey area in the right frame of Figure 6).

At time t_0, p has (x_0, y_0) as coordinates and (x_t, y_t) at time t where

$$x_t = x_{t_0} + \sum_{i=0}^{t} dx_i, \qquad y_t = y_{t_0} + \sum_{i=0}^{t} dy_i \qquad (14)$$

and (dx_i, dy_i) the elementary displacement computed with motion parameters at time i.

Here, as in projection of regions described in Section 4.1, we take the nearest pixel position for each pixel p of Back_0 at time t if the projected coordinates are not integer. (Nevertheless other interpolation methods can be used to compute the projected grey-level value.)

The principle of segmentation correction is based on the assumption of intensity conservation in the background. That is, if the motion is well estimated, then the displaced frame difference (DFD) satisfies

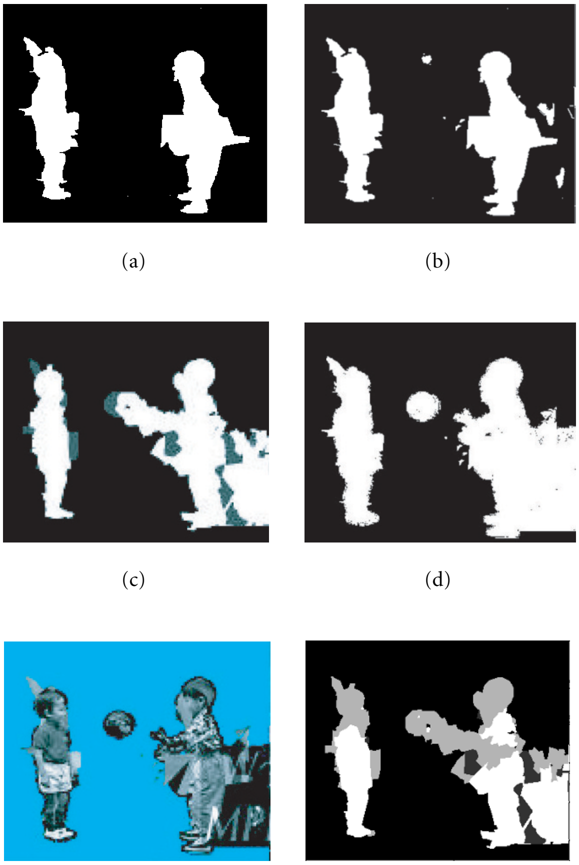
$$\text{DFD}\left(x_t, y_t\right) = I^t\left(x_t, y_t\right) - I^{t_0}\left(x_{t_0}, y_{t_0}\right) = 0, \qquad (15)$$

where (x_t, y_t) are computed by (14).

When we obtain the projection of Back_0 at time t we can reconstruct its grey-value, for that we get back the initial value of each pixel $p \in \text{Back}_0$ stored at initial time t_0. Figure 7b corresponds to the projected Back_0 mask.

5.2. Segmentation correction

The problem now is to project the background Back_0 into the image plane at time t and to compare the projected grey-level value with the current value by means of (15). If new objects appear in the background, then the DFD (15) will be strong. Nevertheless, if the segmentation error is observed at time t due to the false merging of the background and objects, then the low value of DFD (15) in falsely labelled pixels could help the correction. Thus the correction rule is as follows: If $|\text{DFD}(p)| > \text{Th}$ in a pixel corresponding to the projected background, then the pixel p is considered to belong to an object. Otherwise the pixel p belongs to the background.



(a) (b) (c) (d) (e) (f)

FIGURE 7: Extraction of the ambiguous region. (a) Back_0 at t_0, (b) Back_0 at time $t = 15$, (c) the tracking label map at time $t = 15$, (d) the Back_0 label map at time $t = 15$, (e) the DFD frame at time $t = 15$, (f) the ambiguous region in grey.

Figure 7 illustrates this approach. Figure 7a shows the initial state of the segmentation—"certain" background is depicted in black. Figure 7f represents a divergent segmentation with merged object and background. Figure 7b shows the background mask projected at frame at time $t = 13$. Figure 7e shows the difference computed on projected mask of Figure 7b (objects are added for better comprehension). Figure 7d shows the result of such a labelling (the Background is depicted in black). We call the obtained map a *semantic bringing map.* Then the current semantic labelling map (Object, Background, Uncertain) issued from the complete tracking, will be compared with the semantic brining mask. Here we realise a VOP validation and the background validation in the current frame at t.

The background validation

Suppose that a part of the background was falsely merged with a VOP. This yields to ambiguously labelled pixels, which are labelled Object or Uncertain in a tracked segmentation map and Background in the semantic brining mask. To remedy to that we select all those ambiguous regions. Each of

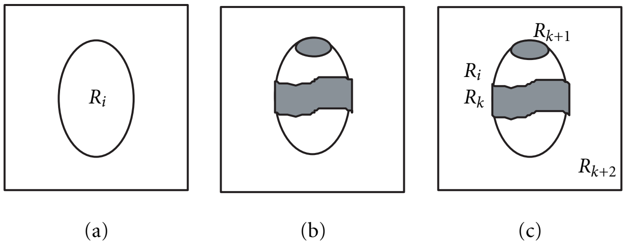


FIGURE 8: The split labelling using the bringing process: (a) tracking result, (b) bringing partition, (c) the correction labelling.

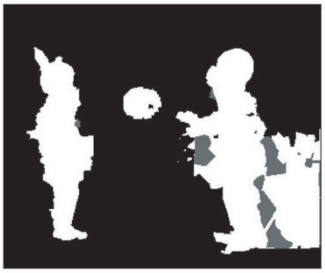


FIGURE 9: The tracking label map after the VOP validation.

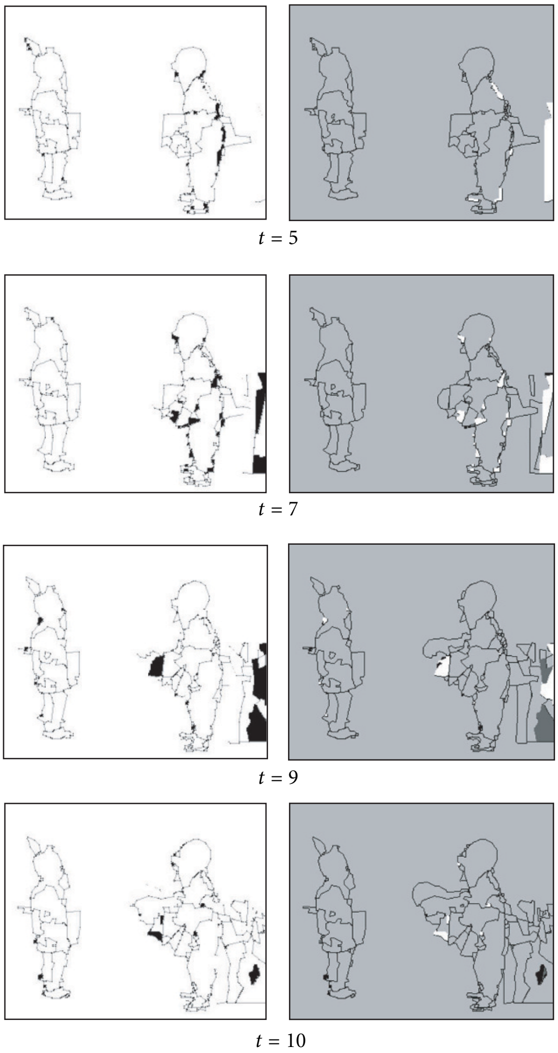


FIGURE 10: Uncovered region classification: the left frames are the uncovered area in black and the right frames are the result of the classification process (sequence "Children"). Background is in black, Object in white, and Uncertain in dark grey. (The pale grey-level in right frames is the color of the rest of region.)

them is re-segmented in one or more new Background regions depending on the distribution of pixels with weak DFD inside. For the rest of pixels they remain in their class (see Figures 8 and 9). Figure 8 helps the reader to focus on the bringing process. Figure 8a displays an ambiguous region R_i resulting from tracking, where Label(R_i) = Object, and label in bringing map (Figure 8b) is Background depicted as dark areas. Then we split R_i into:

- the new Background regions R_k and R_{k+1},
- the rest of pixel set, structured in connected regions, in this case two regions R_i and R_{k+2}.

Figure 9 illustrates the bringing process on a test image.

In order to regularise segmentation map, we realise a preliminary filtering of small connected components in semantic bringing mask.

As an effect of this validation step some or all new regions (uncovered/split) are labelled as Object or as Background. So the following step is the VOP indexing. Each new Object region is affected to a preexisting VOP if it is directly adjacent to the VOP, otherwise it is indexed as a new VOP. The final step in our tracking scheme is the merging process which is realised separately in the VOP and the Background, only the Uncertain regions can be merged indifferently to VOP or Background.

6. RESULTS AND PERSPECTIVES

The proposed automatic indexing of objects in scenes with a changed content was experimented on the sequences "Children," "Akiyo," and "Coastguard" in Common Intermediate Format (CIF) at 12 frames/sec (MPEG4 test sequences).

The quality of obtained segmentation maps was assessed visually in terms of extraction and tracking of principal objects.

The results of semantic classification of uncovered areas are shown in Figure 10. Here the left image depicts these areas in grey, the right image corresponds to the results of classification.

The labelling of *split* regions is illustrated in Figure 11. It can be seen that new moving objects are correctly labelled.

Finally, the tracking results are shown in Figures 12 and 13, the validation level is shown in Figure 14.

Figure 15 depicts the results of two tracking obtained by two closed works: that one presented in [6] and the present one. As it can be seen from the images at Figures 15a and

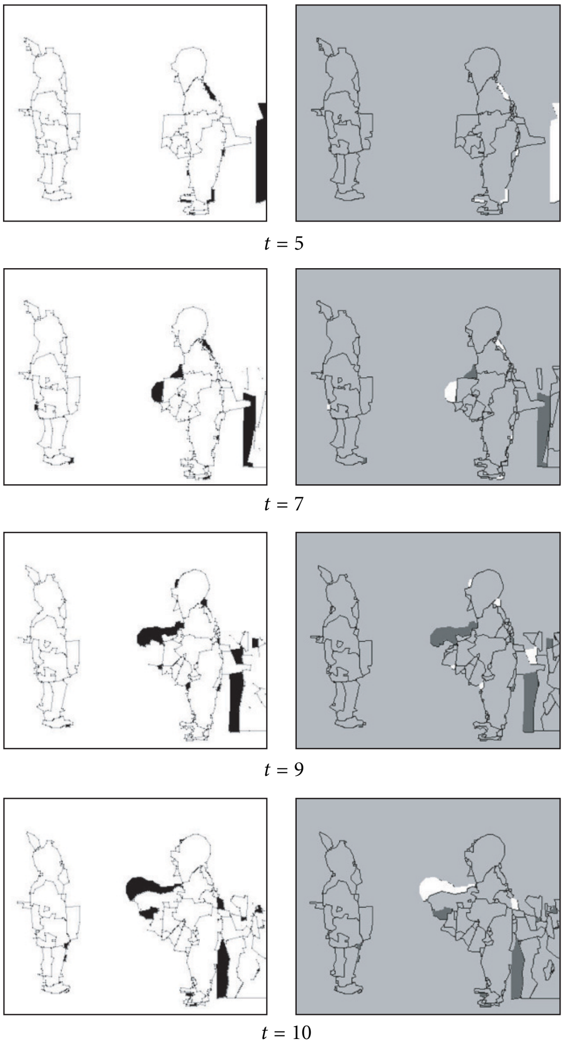


Figure 11: Split region classification: the left frames are the split area in black and the right frames are the result of the classification process (sequence "Children"). Background is in black, Object in white, and Uncertain in dark grey. (The pale grey-level in right frames is the color of the rest of region.)

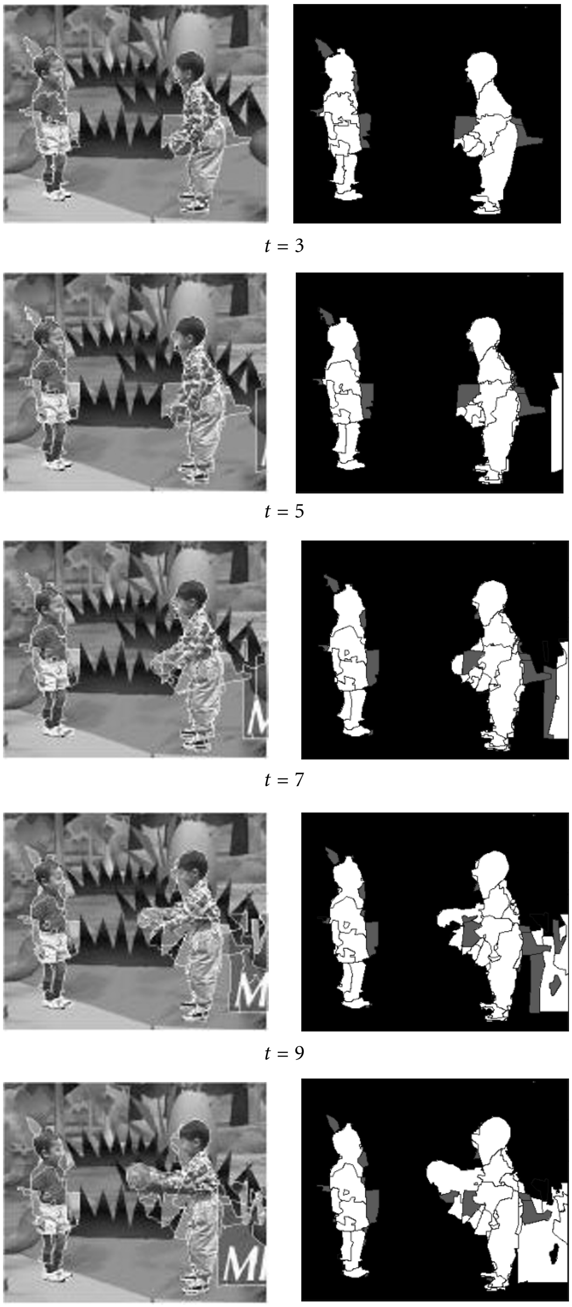
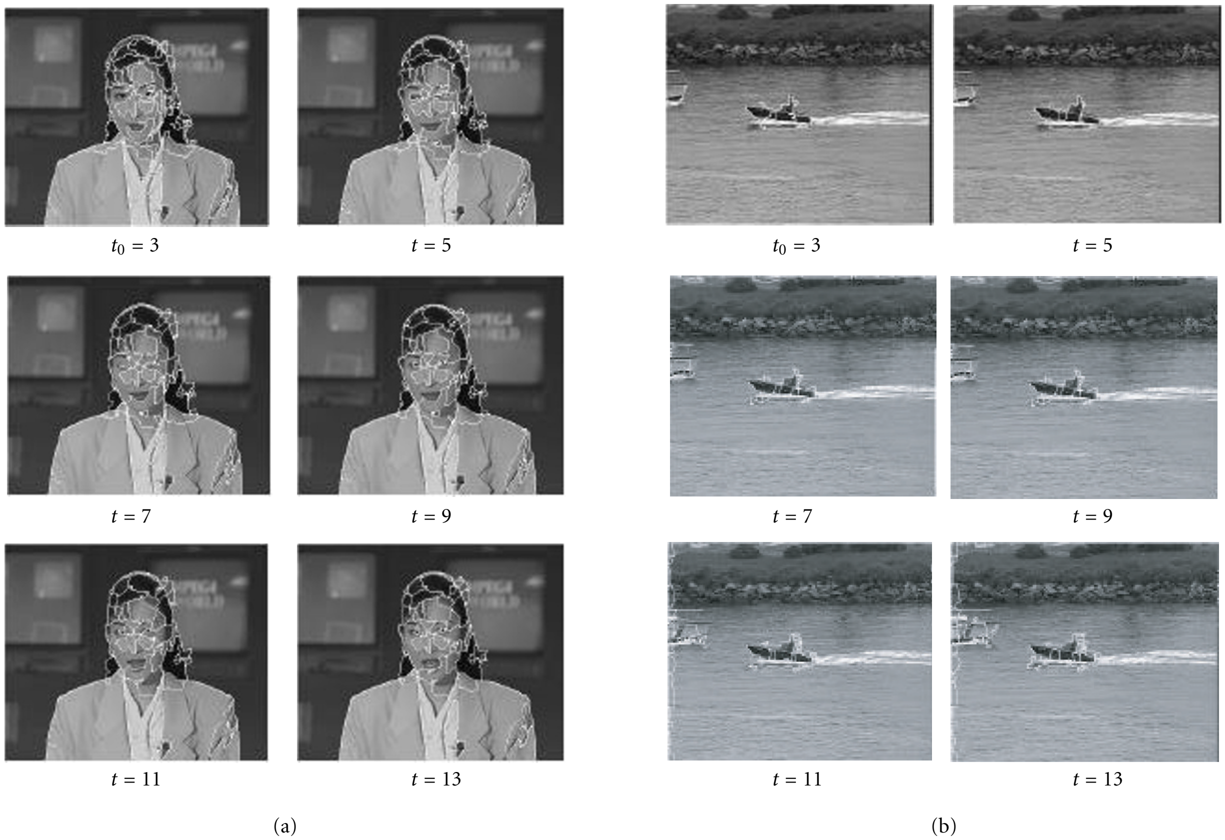


Figure 12: The semantic label of the segmentation. (Sequence "Children.") The left column shows the segmentation, the right column shows the semantic label. ■ Background, □ Object, ▨ Uncertain.

15b at time $t = 5$, the two methods were applied to the same initial partition of the frame into a set of spatio-temporal regions described in Section 2. Figure 15a shows the result of tracking by the method described in [6]. Figure 15b depicts results of tracking method proposed in this paper. Compared to our results, the boundary obtained by the method described in [6] for the same extracted area is more precise. This is due to the use of the spatial (grey-level and colour-based) segmentation. Generally speaking, our method gives an overestimation of object area, while the method used in [6] gives the *interior bound* of objects. What is especially apparent is the capacity of our method to track all meaningful objects in the scene, that is, the objects with sufficiently strong relative motion compared to the motion of the background. (The method in [6] "loses" the ball in the sequence "Children" while our method is able to catch it at any time it moves.) Furthermore, both methods are not free from artefacts when relative motion of regions is weak (see the "handball" region at time $t = 9$). Nevertheless, the forward bringing process of VOP validation corrects the false labelling, and thus shows the strength of the developed method.

These results indicate that in a sequence with a changed content and a strong relative motion of objects with the background the main objects are detected successfully.



$t_0 = 3$ $t = 5$

$t = 7$ $t = 9$

$t = 11$ $t = 13$

(a)

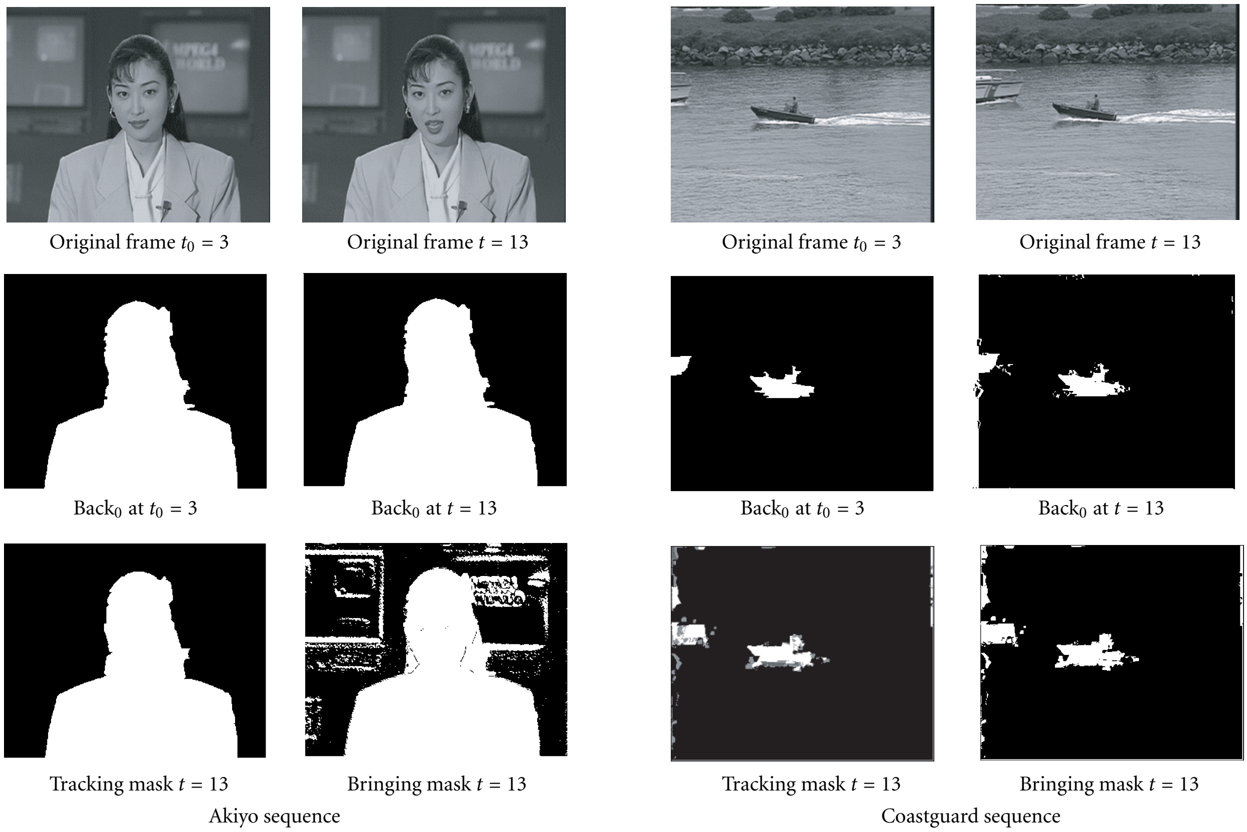
$t_0 = 3$ $t = 5$

$t = 7$ $t = 9$

$t = 11$ $t = 13$

(b)

Figure 13: Tracking result (a) Sequence "Akiyo," (b) sequence "Coastguard."



Original frame $t_0 = 3$ Original frame $t = 13$

Back$_0$ at $t_0 = 3$ Back$_0$ at $t = 13$

Tracking mask $t = 13$ Bringing mask $t = 13$

Akiyo sequence

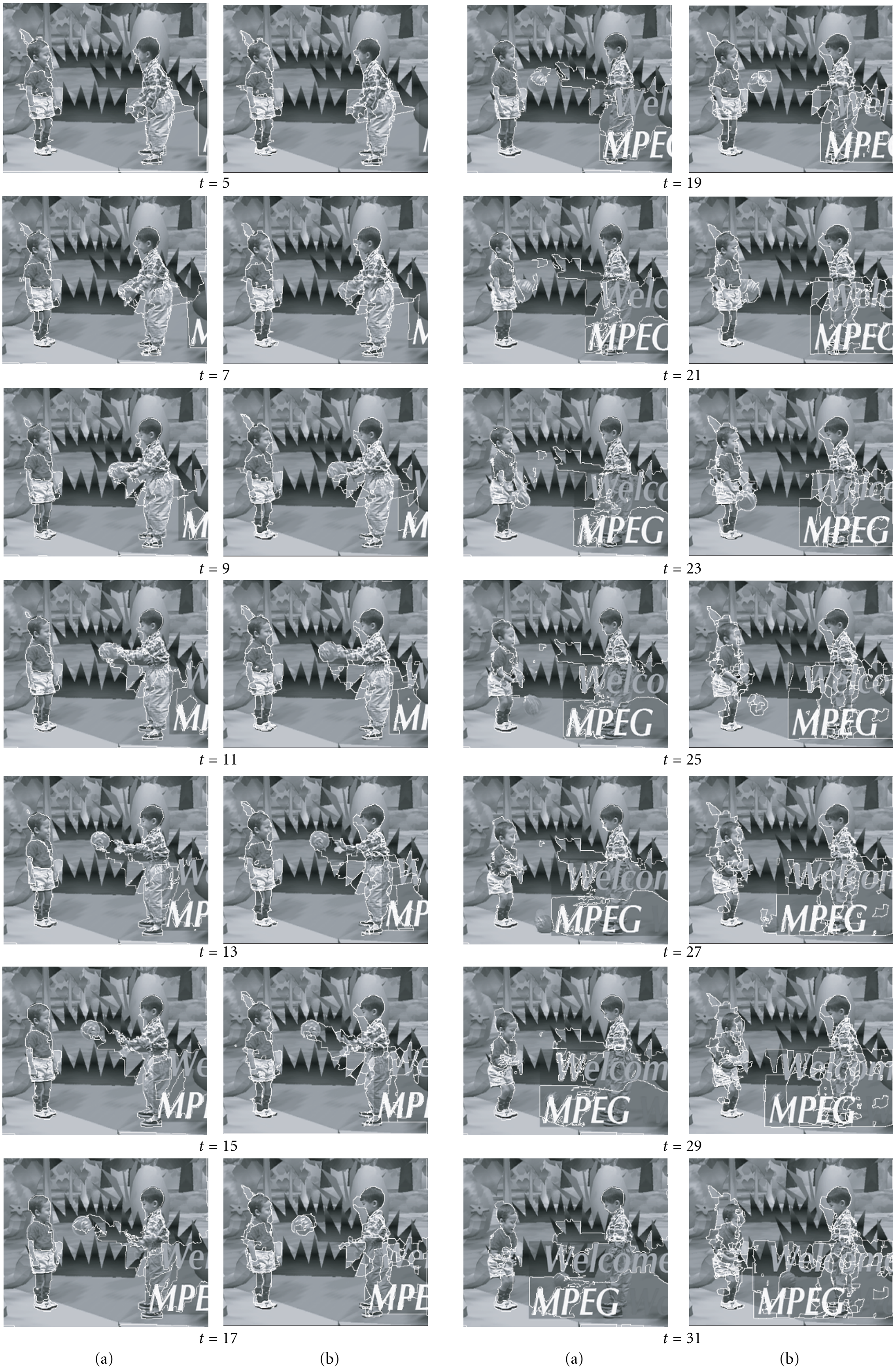
Original frame $t_0 = 3$ Original frame $t = 13$

Back$_0$ at $t_0 = 3$ Back$_0$ at $t = 13$

Tracking mask $t = 13$ Bringing mask $t = 13$

Coastguard sequence

Figure 14: The result of the backward bringing process.



$t = 5$

$t = 7$

$t = 9$

$t = 11$

$t = 13$

$t = 15$

$t = 17$

$t = 19$

$t = 21$

$t = 23$

$t = 25$

$t = 27$

$t = 29$

$t = 31$

(a) (b) (a) (b)

Figure 15: Comparison of tracking results comparison on the sequence Children (a) method [6], (b) proposed method.

Nevertheless, the tracking of the semantic classification can present some errors in thin areas and unavoidable mistakes due to errors of motion estimation. Fortunately, the situation is set upright again by the introduction of the forward bringing process, whose first results are promising. We hope that this method can provide an alternative to human update along tracking.

ACKNOWLEDGMENTS

The authors thank Dr. Stefan Pateux, from INRIA/IRISA, who kindly provided Figure 15a, the tracking results of "Children" sequence using the method described in [6]. This work was supported by the French national RNRT (Réseau National de Recherche en Télécommunication) research project in the frame work of project OSIAM.

REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 N2202, "Information technology-coding of audio-visual objects: visual," ISO/IEC 14496-2 Committee Draft (MPEG4: Visual), Tokyo, March 1998.

[2] F. Marqués and C. Molina, "Object tracking for content-based functionalities," in *SPIE Conference on Visual Communications and Image Processing*, vol. 3024, pp. 190–199, San Jose, Calif, USA, February 1997.

[3] F. Marqués and J. Llach, "Tracking of generic objects for video object generation," in *IEEE International Conference on Image Processing*, vol. 3, pp. 628–632, Chicago, Ill, USA, October 1998.

[4] S. Jehan, M. Barlaud, and G. Aubert, "Detection and tracking of moving objects using a new level set based method," in *International Conference on Pattern Recognition, Signal Processing Conference*, Barcelonne, Spain, September 2000.

[5] P. Frederic and M. Barlaud, "B-spline active contour for fast video segmentation," in *Workshop on Image Analysis for Multimedia Interactive Services*, pp. 47–51, Tampere, Finland, May 2001.

[6] S. Pateux, "Tracking of video objects using a backward projection technique," in *Proc. Visual Communication and Image Processing*, vol. 4067, pp. 1107–1114, Perth, Australia, June 2000.

[7] P. Salembier, "Morphological multiscale segmentation for image coding," *Signal Processing*, vol. 38, no. 3, pp. 359–386, 1994.

[8] A. Mahboubi, J. Benois-Pineau, and D. Barba, "Segmentation spatiale couleur des images par une approche morphologique et hiérarchique," in *CORESA '2000*, Poitiers, France, 19–20 October 2000, session-algorithmes pour fonctionnalités avancées-.

[9] H. Nicolas and C. Labit, "Region-based motion estimation using deterministic relaxation schemes for image sequence coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, pp. 265–268, San Francisco, Calif, USA, March 1992.

[10] L. Wu, J. Benois-Pineau, P. Delagnes, and D. Barba, "Spatio-temporal segmentation of image sequences for object-oriented low bit-rate image coding," *Signal Processing: Image Communication*, vol. 8, no. 6, pp. 513–544, 1996.

[11] J. Benois-Pineau, F. Morier, D. Barba, and H. Sanson, "Hierarchical segmentation of video sequences for content manipulation and adaptive coding," *Signal Processing*, vol. 66, no. 2, pp. 181–201, 1998.

[12] C. Prins, *Algorithmes de Graphes*, Eyrolles, Paris, 1994.

[13] L. Wu, *Segmentation spatio-temporelle d'images animées en vue d'un codage à fort taux de compression*, Ph.D. thesis, université de Nantes, France, 1995.

[14] J. Benois-Pineau, L. Wu, and D. Barba, "Content-based border preserving coding for structure retrieval and content manipulation of image sequences," in *Picture Coding Symposium Advance Program*, pp. 553–558, Melbourne, Australia, March 1996.

[15] F. Heitz and P. Bouthemy, "Multimodal motion estimation and segmentation using Markov random fields," in *Proc. 10th IEEE International Conference Pattern Recognition*, vol. 1, pp. 378–383, Atlantic City, NJ, USA, June 1990.

[16] H. Nicolas, "Mosaic representation and video object manipulations for post-production applications," in *IEEE Int. Conf. on Image Processing*, vol. 2, pp. 451–455, Chigaco, Ill, USA, October 1998.

[17] S. Peleg, B. Rousso, A. Rav-Acha, and A. Zomet, "Mosaicing on adaptive manifolds," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1144–1154, 2000.

[18] H. Walin, C. Christopoulos, A. Smolic, Y. Abdeljaoued, and T. Ebrahimi, "Robust mosaic construction algorithm," ISO/IEC JTC1/SC29/WG11 MPEG00/M5698, Report, Noordwijkerhout, The Netherlands, March 2000.

[19] L. Bonnaud, C. Labit, and J. Konrad, "Interpolative coding of image sequences using temporal linking of motion-based segmentation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 2265–2268, Detroit, Mich, USA, May 1995.

Amal Mahboubi was born in Sidi-Bel-Abbes, Algeria, on February 11, 1972. She received the engineering degree in computer science from Sidi-Bel-Abbes University in 1997 and DEA degree in 1999 from Paul Sabatier University, Toulouse, France. In 1999, she joined the IRCCyN laboratory as a doctoral student, she is currently working toward the Ph.D. degree in applied automatic computer science. Her current research interests are in the areas of video analysis.

Jenny Benois-Pineau graduated from Moscow Technical University of Electronic Engineering (MIET, Moscow, Russia) and received her Ph.D. degree in computer science and control systems in 1989 from this university. In 1990 she joined l'Ecole Polytechnique de l'Université de Nantes for her post-doctoral research in image processing and coding. Since 1992 she has worked as associate professor at EPUN. In September 2001 she got a full professor position in computer science at the University of Bordeaux 1 UMR CNRS LABRI where she is doing her research in multimedia and more precisely in video analysis, indexing, coding, and image processing. Professor Benois is the author and coauthor of more than 60 journal papers, conference proceedings, invited lectures. She worked as invited lecturer at the University of Sussex (GB), University of Växjo (Sweden), Technical University of Malaysia, Technical University of Budapest. She was a leading researcher in French-Greek and French-Hungarian research programs on model-based and object based video coding. She has been involved into large French research programs on video segmentation and coding. Her research interests include multimedia signal processing and video analysis. Professor Benois is a member of IEEE.

Dominique Barba was born in France on June 6, 1944. He recieved the Ph.D. in telecommunications (University of Rennes) and doctorate in mathematical sciences, speciality computer sciences (University of Paris VI) in 1972, 1981, respectively. From 1973 to 1984, he was Assistant Professor at INSA of Rennes, working in the field of digital image processing with psychovisual quality criterium. From 1984 to 1985, he was attached to the CNET/CCETT research centre in charge of image and video compression. Since 1986, he moved as a full Professor to Polytech'Nantes (Ex IRESTE Institute of University of Nantes), and created a laboratory in Digital Image Processing which has become a part of IRCCyN.