pp. 408-415

Allowing content-based functionalities in segmentation-based coding schemes

Beatriz MARCOTEGUI *
Ferran MARQUÉS **
Fernand MEYER *

Abstract

This paper deals with the use of the segmentation tools and principles presented in [10] and [13] for allowing content-based functionalities. In this framework, means for supervised selection of objects in the scene are proposed. In addition, a technique for object tracking in the context of segmentation-based video coding is presented. The technique is independent of the type of segmentation approach used in the coding scheme. The algorithm relies on a double partition of the image that yields spatially homogeneous regions. This double partition permits to obtain the position and shape of the previous object in the current image while computing the projected partition. In order to demonstrate the potentialities of this algorithm, it is applied in a specific coding scheme so that content-based functionalities, such as selective coding, are allowed.

Key words : Image processing, Segmentation, Image coding, Moving image, Tracking, Interactive system.

FONCTIONNALITÉS LIÉES AU CONTENU DANS LES SYSTÈMES DE CODAGE À BASE DE SEGMENTATION

Résumé

Cet article traite de l'utilisation d'outils de segmentation permettant la mise en place de fonctionnalités liées au contenu. Dans ce cadre, des techniques de sélection supervisée d'objets sont proposées. De plus, une technique de suivi d'objet est présentée. Cette technique est indépendante du type de segmentation utilisée pour le système de codage. L'algorithme repose sur une double partition conduisant à des régions spatialement homogènes. Cette double partition permet d'obtenir la position et la forme des objets présents dans la trame précédente. L'intérêt de l'algorithme est illustré dans le contexte d'un système de codage sélectif.

Mots clés : Traitement image, Segmentation, Codage image, Image animée, Poursuite, Système interactif.

Contents

I. *Introduction.*
II. *Definition of objects.*
III. *Tracking of objects.*
IV. *An example of content-based functionality : selective coding.*
V. *Conclusions.*
References (21 ref.).

I. INTRODUCTION

In the framework of video coding, new coding schemes allowing content-based functionalities are a very active research field [3]. Video coding algorithms with embedded content-based functionalities enable the separate manipulation and definition of the various objects in the scene. These features permit, for instance, content-based multimedia data access or content-based scalability.

* Centre de Morphologie mathématique, 35, rue Saint-Honoré, F-77305 Fontainebleau Cedex, France. marcotegui@cmm.ensmp.fr. meyer@cmm.ensmp.fr.
** UPC Campus Nord, Modulo D5, c. Gran Capita, s/n, 08034 Barcelona, Espagne. ferran@gps.tsc.upc.es.

These concepts are driving the MPEG-4 standardization effort. The mandate of MPEG-4 is to define a new coding standard, allowing new ways of communication, manipulation and access of digital audio-visual data [17]. MPEG-4 addresses these tasks from the point of view of integration. That is, MPEG-4 tries to integrate natural and synthetic audio-visual information as well as different kinds of representation such as 2D and 3D models, stereo and multiview video.

The definition of content-based functionalities on this integrated representation demands a description of image sequences in terms of objects or groups of objects. Separate objects can be automatically obtained in the recording process by means, for example, of a chroma-keying technique. Figure 1 shows an example of an image created by the composition of two objects that have been separately recorded.

However, to be able to apply content-based functionalities to any kind of video data, techniques for defining the set of objects present in a scene are necessary. This definition can be done either automatically (in an unsupervised way) or with the help of the user. Supervised object definition can even be done from the receiver side, in the case of interactive applications.

Following the discussion in [10], two strategies may be applied for analysing the content of an image sequence. Classical object definition techniques rely on motion information to divide the image into objects [1, 2, 4]. The unsupervised algorithm will detect a series of moving objects; user interaction will be reduced to the selection of one of those. Hence only moving objects may be selected, and the remaining part of the information is out of the scope of interaction. New approaches are necessary in order to allow the user to select parts of the moving object or even static objects on which functionalities are to be applied.

User's interaction should be a simple as possible. The definition of an object should not demand the user to perfectly define the object boundaries. On the contrary, a sketch of the object shape to be tracked should be sufficient to define the object. For instance, if the selected object is the face of a person, simple marks like a circle or a cross should be enough information for the algorithm to obtain the shape of the face.

Once objects have been defined, a tracking technique should be able to follow them along the sequence. It is this tracking capability which really opens the door to content-based functionalities. In the unsupervised case, it allows to relate the information of the objects in the previous frames to that of the current and future frames. This leads to the creation of video object representations; that is, to have separately the temporal evolution of an object. In addition to that, in the supervised case, it allows the user to mark only once the selected object.

Classical object tracking techniques [21, 20] do not completely fulfill the requirements of this application. First, since they rely on motion information, parts of an object with homogeneous motion or a combination of objects with different motions cannot be correctly tracked. In addition to that, global camera motion and still (or stopping) object may lead to tracking problems.

Second, classical object tracking techniques do not interact with coding algorithms. In the framework of video coding functionalities, zones to be tracked should not only correspond to the objects in the scene, but also they should be amenable to an efficient coding. In addition, tracking algorithms have to be able to cope with special constraints of the coding algorithm. For instance, they should not lose track of the objects when a frame in the sequence is coded in intra-mode.

Finally, object-oriented video coding schemes base their segmentation on a manifold of homogeneity criteria [10] (*e.g.* : motion homogeneity, gray level homogeneity, special type of texture homogeneity, etc) or even a combination of some criteria, as it has been seen in [13]. A new type of objects will have to be added : user defined objects. Object tracking techniques for content-based functionalities will have to follow simultaneously the regions required for an efficient coding and the user defined objects. Therefore, they have to cope with regions that may be defined following very different homogeneity criteria.

This work presents how the segmentation techniques presented in previous sections have been adapted to allow content-based functionalities in video coding algorithms. We will show successively how the tools presented in [13] are used as basis for object definition and object tracking algorithms.

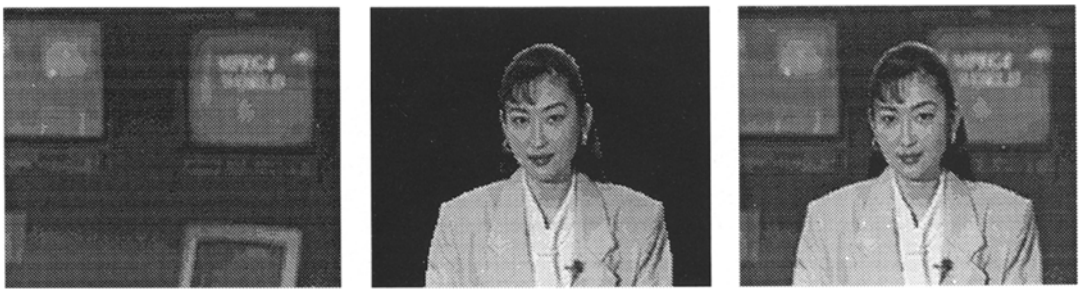

FIG. 1. — Background and foreground objects from frame #0 in *Akiyo* sequence and their composition.

Objets du premier plan et de l'arrière-plan pour la trame #0 de la séquence Akiyo *et leur composition.*

User's interaction can be easily introduced in these segmentation schemes by means of the concept of markers. Simple sketches can be utilized as external markers in the segmentation procedure or being used to validate regions that have been already obtained.

In order to have an object tracking technique independent of the type of segmentation criteria used in the coding approach, a double segmentation is applied, using the watershed algorithm [12]. We already met the necessity for a double segmentation in [7] when motion homogeneity merged a large number of regions with different textures; in this case, the time projection of regions which are a patchwork of various textures does not work anymore. A fine partition has to be built in parallel with the actual partition, in order to project regions with homogeneous textures. In the case of object tracking, we introduce a supplementary constraint : the contours of the object of interest have to be always present in both partitions.

This paper is structured as follows. In Section II, we describe how user interaction may be built in the segmentation process, in order to select and define objects of interest. Section III is devoted to the description of a technique for tracking objects through the sequence. This method can be seen as an extension of the *Projection* step presented in [16]. To validate this technique, Section IV describes the way to introduce it in a specific segmentation-based coding scheme so that content-based functionalities are addressed. This coding scheme relies on the concepts of *Partition tree* and *decision tree* described in [6]. Finally, Section V presents some conclusions as well as future work.

II. DEFINITION OF OBJECTS

One of the major points of morphological segmentation is that it separates the problem of selection of zones of interest from the problem of final location of the contours related to those zones. Therefore, it allows the introduction of external markers in the *decision* procedure. Such markers can be added to those obtained by the marker extraction procedure [13], so that the final partition contains new regions associated to the user selected markers.

A second possibility for the definition of objects is a direct interaction of the user with a given partition of the image. In this case, the marks introduced by the user are not used as markers in the segmentation procedure but they are related to a set of regions in the partition.

In both cases, a first approximation of the desired object is obtained and the final shape of the object can be further improved by the user following the same strategy. These refinements may be necessary if, for instance, the marks are not representative enough to obtain a good region, in the first approach; or the set of regions forming the initial partition do not correctly correspond to the selected object, in the second approach.

An example of the selection of an object from a rough mark is presented in Figure 2. In the first image, a mark prompting the child in the first frame of the sequence *mother and daughter* is superposed to the original image. The second image presents the selected object obtained using this mark.

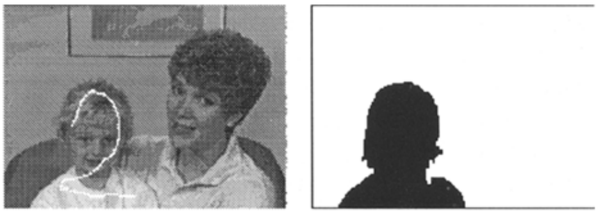

FIG. 2. — Selection of an area of interest from the sequence *mother and daughter*.

Sélection d'une zone d'intérêt de la séquence mother and daughter.

III. TRACKING OF OBJECTS

Usual segmentation-based coding approaches utilize motion or/and gray level homogeneity in order to compute the image partition for coding purposes. The contours which are produced do not necessarily follow the boundaries of the selected object. Let's illutrate this concept with an example. In Figure 3, three frames of the sequence *foreman* are shown. Let us asume that the

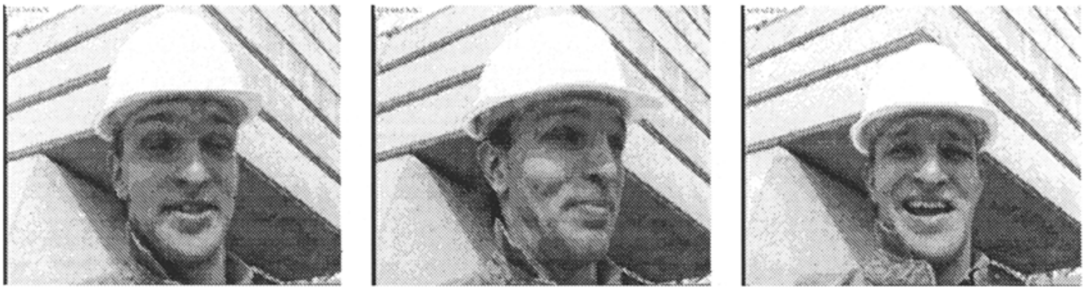

FIG. 3. — Original frames number 0, 5 and 20 from the sequence *foreman*.

Trames originales #0, #5 et #20 de la séquence foreman.

object marked by the user is the head of the person. In some frames, a segmentation technique relying on gray level criteria may merge the helmet with the building, since they have very close gray level values. In turn, if the segmentation is based on motion information, some zones of the background may be jointly segmented with some regions of the face, in case of global motion of the camera, or some parts of the body of the person may be merged to the face, if the man moves like a rigid object.

Therefore, in order to track the regions that formed the partition in the previous frame, the algorithm (that is, the *projection* step) may have to handle regions with different homogeneity nature. In the work presented in [14, 5], the regions detected in frame $t - 1$ are first motion compensated. They are then used as markers and extended by a 3D watershed algorithm in the current image, assuming that regions are spatially homogeneous. However, this asumption is wrong in the present case. To solve this problem, a tracking technique relying on a double partition approach is proposed [9].

In this approach, two different levels of partition are defined. The partition of the previous image P_{T-1} is re-segmented in order to achieve a finer partition. This fine partition contains a larger number of regions which are obtained by re-segmenting the regions in P_{T-1} following spatial criteria. Spatially homogeneous regions are created; being sub-regions of the coarse partition, the contours of the coarse partition, and in particular the contours of the objects of interest are also present in the fine partition.

III.1. Creation of the fine partition.

Let us assume that frame T in a sequence has been already coded and, therefore, its partition is available. This partition should contain a region or a set of regions that defines the object to be tracked. The first step to obtain the object in frame $(T+1)$ is to produce a fine partition of the previously coded frame T. This fine partition is obtained by splitting the coded coarse partition into several regions. The segmentation procedure is constrained by the coarse partition already obtained for frame T in the coding procedure. This is done by using a constrained watershed algorithm [18].

The process of creating this fine partition is purely intra-mode and, therefore, no motion information is used at this stage. The homogeneity criteria that are used only deal with size and contrast information [18]. This fine partition is created in order to make the projection of the previous partition easier and, therefore, the tracking of the object. This procedure allows obtaining a partition with good features for object tracking purposes, independently of the type of partition used at the coding level.

III.2. Projection of the finest partition.

This fine partition is then projected in the current frame in order to obtain a fine segmentation at time T. Several techniques can be used in order to carry out this fine partition projection [15, 5, 11]. In this work, the algorithm presented in [15] has been modified in order to improve the temporal coherence of the regions through the sequence partition. In [15], the watershed algorithm [12] was extended to the case of image sequences. In addition, the cost of assigning a pixel p_i to a region r_j was modified to take into account the contour complexity of the resulting regions. Here, a new cost function is proposed that controls the deformation of the marker. It prevents projected regions to grow too far from the previous markers [8]. Therefore, the cost of assigning a pixel p_i to a region r_j uses three different types of information :

$$(1) \quad cost(p_i, r_j) = \alpha_1 dist_t(p_i, r_j) + \alpha_2 dist_c(p_i, r_j) + \alpha_3 dist_d(p_i, r_j).$$

The three functions $dist_t$, $dist_c$ and $dist_d$ are the distances related to the texture, contour complexity and deformation information, respectively. The function $dist_t$ computes the difference between the mean value of a region r_j and the gray level value of a neighboring p_i. In turn, the function $dist_c$ is related to the increase in contour complexity of a region r_j as a pixel p_i is added to it. Finally, the function $dist_d$ measures the deformation of a region r_j with respect to the projected marker when adding a pixel p_i to it [8].

The use of this cost function increases the stability of the labels through the time domain and, therefore, allows a better tracking of the objects.

III.3. Obtaining the projected area of interest.

In order to obtain the partition P_T, only those contours of the fine projected partition associated to the coarse partition have to be kept. This can be easily done since the projection algorithm keeps track of the labels of each region. Therefore, the projection of a region from the coarse partition can be obtained by merging the projections of the regions that belong to it at time T; that is, by re-labeling the fine projected partition. However, the projection and re-labeling of the fine partition does not ensure that an actual partition is obtained.

Indeed, unconnected regions may have the same label after relabeling. According the type of applications, this may or may not be a problem. If an object is allowed to split in several part, nothing more has to be done. If one desires keeping a single connected component through the sequence, one keeps at each step only the largest component for each label and removes the other components with the same label. This creates some holes or uncertainty zones, easily assigned to neighboring regions by a pure 2D watershed algorithm. The complete procedure is illustrated in Figure 4, where the evolution of a selected object is shown.

III.4. Frames in intra-mode.

In object based coding, the projection plays an important role and is the key-feature for taking advantage of

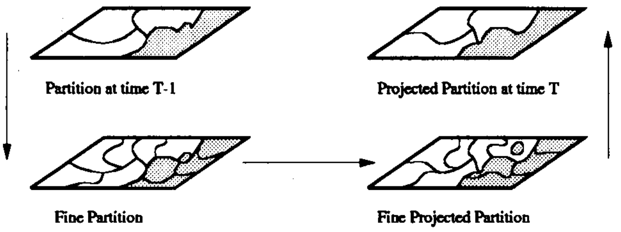

FIG. 4. — Example of tracking of an area of interest.

Exemple de suivi de la zone d'intérêt.

the temporal redundancy in image sequences. After a first frame coded in *intra-mode*, the subsequent frames are coded in *inter-mode*, which mean : prediction from the preceding frame and coding of the errors. However, once in a while, the process has to be refreshed and a new frame coded in *intra-mode*. In this situation, the information on the object to be tracked has to be transmitted from the last coded frame coded in *inter-mode* into the new frame coded in *intra*.

The previous technique can also be applied in the case of using an intra-frame mode to code the current frame. The selected object from frame $T-1$ is projected into frame T. Then, the final partition for frame T is computed in a pure *intra-mode*. However, the final partition is constrained to contain the region associated to the projected object. In this way, the frame T can be coded in *intra-mode* without loosing track of the selected region.

III.5. Results.

In this section, some examples of results obtained with this technique are presented. In Figure 5, the result achieved in the case presented in the introduction is shown; that is, the tracking of the head of the man in the sequence *foreman*. Results for frames 0, 5 and 20 as well as 80, 105 and 185 are presented. These results correspond with some of the frames with strongest motion in the sequence. Note that the algorithm is able to correctly track the head of the man, even if it is formed by two very different objects : the face and the helmet.

In the example of Figure 6, the selected object is the dancer that appears in the TV set of the background. This is a difficult object to be tracked by motion-driven techniques, since the motion of the dancer is clearly not homogeneous. Nevertheless, the object tracking technique previously proposed yields a correct tracking of the dancer.

IV. AN EXAMPLE OF CONTENT-BASED FUNCTIONALITY : SELECTIVE CODING

The above technique for object tracking has been introduced and tested in the complete segmentation scheme presented in [16]. This segmentation scheme, that was originally designed for coding efficiency, is here adapted to address content-based selective coding. This way, given an initial selection of an object, the coding algorithm should be able to track and to code it with better quality than the rest of the image.

IV.1. Modifications to allow content-based selective coding.

In order to address this functionality, the basic blocks of the segmentation scheme have to be modified. In this section, the necessary modifications to allow content-based selective coding are presented. The main modifications are the following ones :

- **Projection :** the partition representing the previous frame contains a region or a group of regions representing the object of interest. In order to obtain the partition for the current frame, we apply the double partition approach presented in [16].
- **Partition tree :** the set of partitions that forms the *partition tree* should be created having in mind the constraint introduced by the projected object of interest. The different proposals of regions contained in the *partition tree* must comply with the task of object tracking. Fusions of regions are specially dangerous with this respect : the merging of regions with similar motion has to be forbidden if they do not belong to the same object. Such a merging would make impossible the separate tracking of these objects.
- **Decision :** this block should yield coding strategies leading to a lower distorsion within the selected object than in the other areas of the image. In order to obtain this selective coding, the bit allocation strategy used in the basic segmentation algorithm is slightly modified. In this case, if a target bit rate has to be reached, selective coding is implemented by giving a higher weight to the distorsion within the object of interest.
- **Coding :** this block does not need to be changed since the coding techniques used in this functionality are the same as those used in the general case.

IV.2. Results.

Figure 7 presents the results of applying the previous algorithm to the sequence *mother and daughter*. The first row in Figure 7 presents three original frames of the sequences, whereas the second row shows the decoded frames. In this example, the head of the mother has been selected as an object to be tracked and selectively coded. The selective coding has been carried out by multiplying by a factor 10 the distorsion inside this area of interest. The whole sequence has been coded at 30 kbit/s and 5 frames/s.

The different quality obtained for the heads of both person in the scene should be highlighted. In addition,

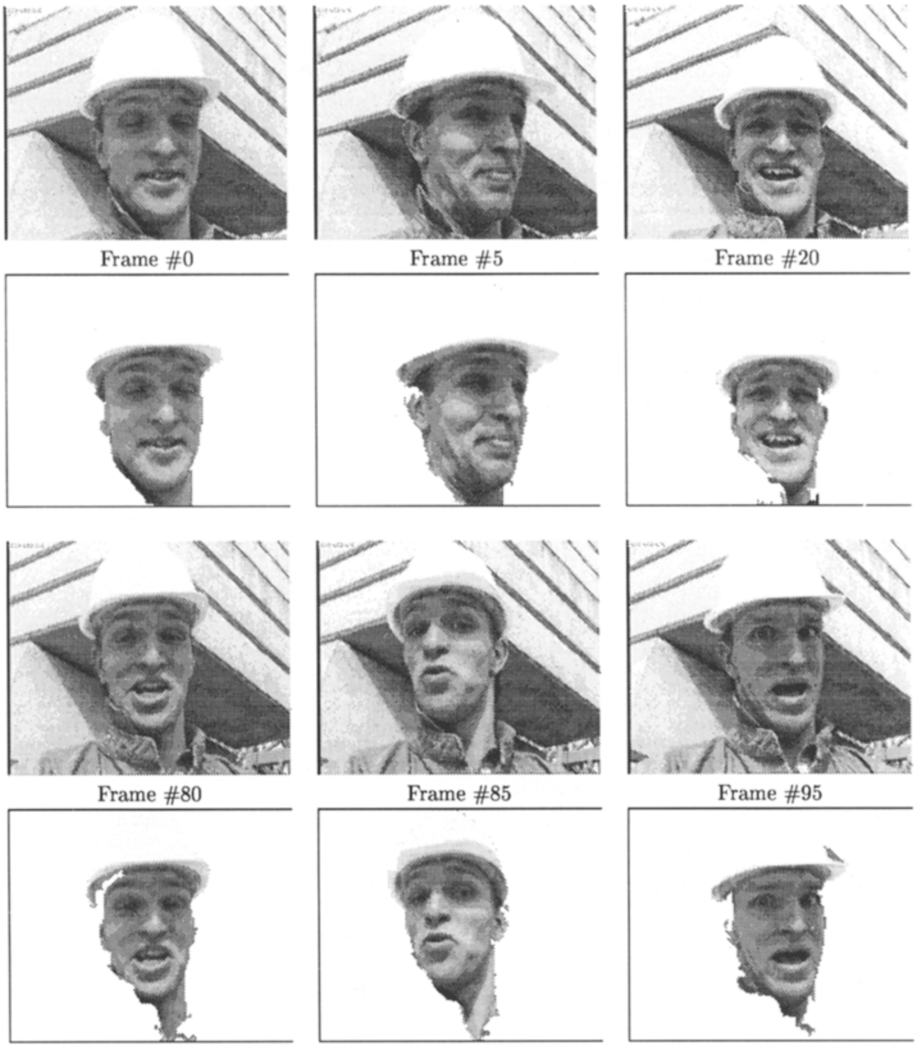

FIG. 5. — The tracking of the head of the man in the sequence *foreman*.

Suivi de la tête du personnage de la séquence foreman.

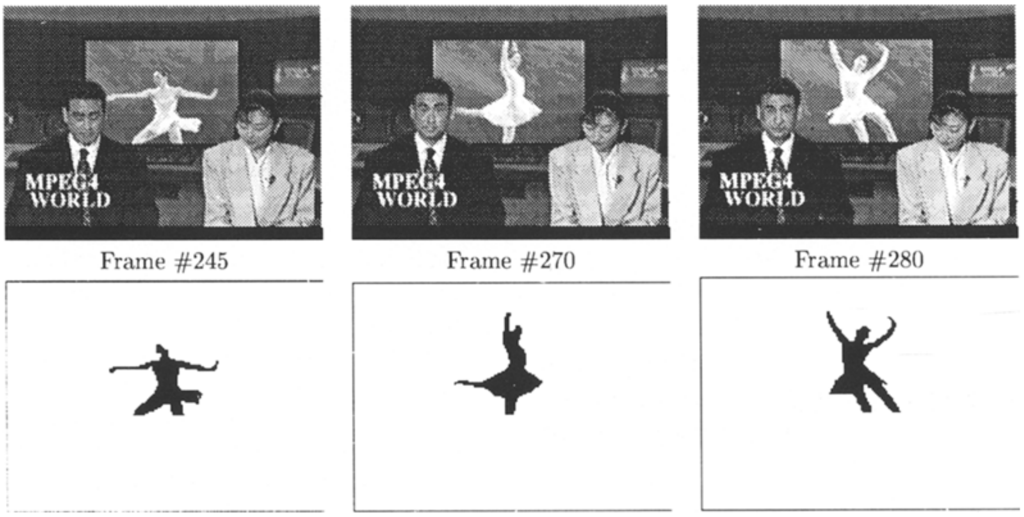

FIG. 6. — The tracking of the dancer in the sequence *news*.

Suivi de la danseuse dans la séquence news.

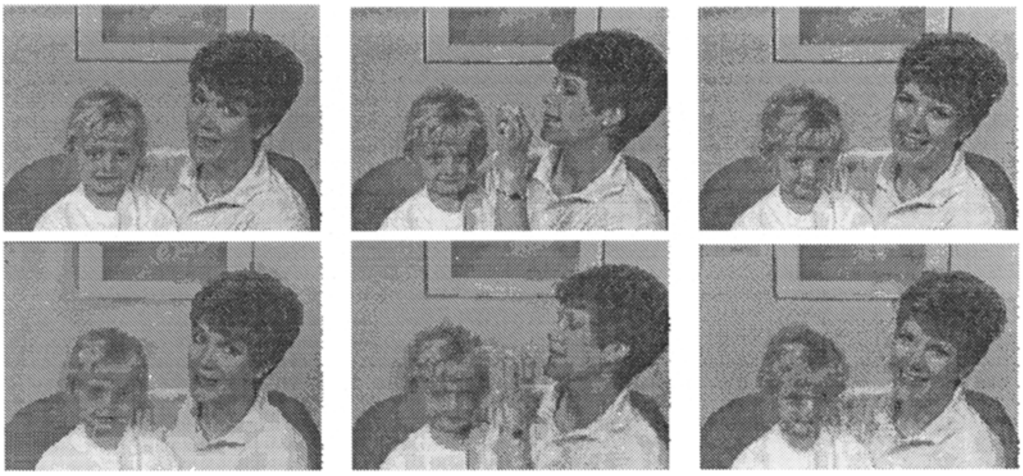

FIG. 7. — The tracking and selective coding of the head of the mother in frames number 0, 84 and 246 of the sequence *mother and daughter.*

Suivi et codage sélectif de la tête de la mère de la séquence mother and daughter, *trames #0, #84 et #246.*

note that the evolution of the mother's face is correctly tracked. This tracking is done in spite of the fact that, between frames 150 and 200, the mother moves up and down her head which results in a partial occlusion of her face. The presence of noise in the decoded image in the zones around the mother's face are due to the fact that these areas are related to the background which is uncovered by the movement of the mother's head.

V. CONCLUSIONS

In this paper, a technique for tracking areas of interest in a segmentation-based video coding scheme has been proposed. The technique does not assume any type of specific partition and, therefore, it can also be applied to block-based coding schemes that use additional contour information such as masks or alpha planes.

In addition, the algorithm gives the possibility to define or refine the set of regions forming an area of interest. Since the segmentation involved in the tracking of areas of interest relies on the concept of markers (that is, uses the watershed algorithm), user's interaction is feasible. Users can mark the areas of interest or refine the existing ones simply by roughly drawing dots or lines inside or around the areas of interest. Such draws can be introduced as markers of regions in the watershed algorithm so that the areas of interest are defined or improved [19].

Once a technique for tracking areas of interest is available, several content-based functionalities can be addressed. In this paper, the basic Sesame coding scheme [19] has been modified in order to enable content-based selective coding. The necessary modifications for such a functionality only involve, in addition to the application of the algorithm for tracking areas of interest previously presented, slight variations of the *partition tree* and *decision* blocks.

The current work aims at the extension of this algorithm so that functionalities such as content-based scalability, both in the time and the space domains, can be addressed.

Manuscrit reçu le 22 mai 1997.

REFERENCES

[1] BOUTHEMY (P.), FRANÇOIS (E.). Motion segmentation and qualitative dynamic scene analysis from an image sequence. *International Journal of Computer Vision* (1993), **10,** n° 2, pp. 157-182.

[2] ***. MPEG-4 Video Group. MPEG-4 automatic segmentation of moving objects (core experiment n2). *In Doc. ISO/IEC JTC1/ SC29/WG11 MPEG96/841* (March 1996).

[3] ***. MPEG-4 proposal package description (PPD). *ISO/IEC JTC1/ SC29/WG11* (July 1995).

[4] KRUSE (S.), KAUFF (P.). Fine segmentation of image objects by means of active contour models using information derived from morphological transformation. *In Visual Communication and Image Processing,* Orlando, USA (Apr. 1996), pp. 1164-1172.

[5] MARCOTEGUI (B.). Segmentation de séquences d'images en vue de codage. *PhD Thesis,* Ecole des Mines de Paris, Fr. (1996).

[6] MARCOTEGUI (B.), MARQUÉS (F.), MORROS (R.), PARDÀS (M.), SALEMBIER (P.). Segmentation of video sequences and rate control. *Ann. Télécommunic.* (1997), **52,** n° 7-8, pp. 380-389.

[7] MARCOTEGUI (B.), MEYER (F.). Bottom-up segmentation of image sequences for coding. *Ann. Télécommunic.* (1997), **52,** n° 7-8, pp. 397-407.

[8] MARQUÉS (F.). Motion stability in image sequence segmentation using the watershed algorithm. In P. Maragos, R. Schafer and M. Butt, editors. *Mathematical morphology and its applications to image and signal processing,* Kluwer *Academic Publishers* (May 1996), pp. 321-328.

[9] MARQUÉS (F.), MARCOTEGUI (B.), MEYER (F.). Tracking areas of interest for content-based functionalities in segmentation-based video coding. *In International Conference on Acoustics, Speech and Signal Processing, ICASSP'96,* Atlanta, USA (May 1996), pp. II.1224-II.1227.

[10] MARQUÉS (F.), MEYER (F.), PARDÀS (M.), SALEMBIER (P.). General requirements for coding oriented segmentation of video-sequences. *Ann. Télécommunic.* (1997), **52,** n° 7-8, pp. 359-366.

[11] MARQUÉS (F.), PARDÀS (M.), SALEMBIER (P.). Coding-oriented segmentation of video sequences. In L. Torres and M. Kunt, editors, *Video Coding : The second generation approach*, Kluwer *Academic Publishers* (1996), pp. 79-124.

[12] MEYER (F.), BEUCHER (S.). Morphological segmentation. *Journal of Visual Communication and Image Representation* (Sep. 1990), **1**, n° 1, pp. 21-46.

[13] MEYER (F.), OLIVERAS (A.), SALEMBIER (P.), VACHIER (C.). Morphological tools for segmentation : connected operators and watersheds. *Ann. Télécommunic.* (1997), **52**, n° 7-8, pp. 367-379.

[14] PARDÀS (M.), SALEMBIER (P.). Joint region and motion estimation with morphological tools. In J. Serra and P. Soille, editors, *Second Workshop on Mathematical Morphology and its Applications to Signal Processing*, Fontainebleau, Fr., Kluwer *Academic Press* (Sep. 1994), pp. 93-100.

[15] PARDÀS (M.), SALEMBIER (P.). Time-recursive segmentation of image sequences. In EURASIP, editor, *EUSIPCO 94, VII European Signal Processing Conference*, Edinburgh, UK (Sep. 13-16, 1994), pp. 18-21.

[16] PARDÀS (M.), SALEMBIER (P.). Segmentation of video sequences for partition tree generation. *Ann. Télécommunic.* (1997), **52**, n° 7-8, pp. 389-396.

[17] PEREIRA (F.). MPEG-4 : a new challenge for the representation of audio-visual information. *In Picture Coding Symposium* (May 1996).

[18] SALEMBIER (P.). Morphologial multiscale segmentation for image coding. *EURASIP Signal Processing* (Sep. 1994), **38**, n° 3, pp. 359-386.

[19] SALEMBIER (P.), MARQUÉS (F.), PARDÀS (M.), MORROS (R.), CORSET (I.), JEANNIN (S.), BOUCHARD (L.), MEYER (F.), MARCOTEGUI (B.). Segmentation-based video coding system allowing the manipulation of objects. *IEEE Trans. on Circuits and Systems for Video Technology* (Feb. 1997), **7**, n° 1, pp. 60-74.

[20] SEZAN (I.), LAGENDIJK (R. L.). Motion analysis and image sequence processing. Kluwer *Academic Publishers*, Boston (1993).

[21] TZIRITAS (G.), LABIT (C.). Motion analysis for image sequence coding. *Elsevier Science B. V.*, The Netherlands (1994).