SEGMENTATION AND TRACKING OF VIDEO OBJECTS FOR A CONTENT-BASED VIDEO INDEXING CONTEXT

Magali Mazière[1], *Françoise Chassaing*[1], *Luis Garrido*[2] *and Philippe Salembier*[2]

[1] France Telecom CNET / DIH / HDM
4, rue du Clos Courtel
35512 Cesson-Sévigné
FRANCE

[2] Universitat Politècnica de Catalunya, Campus Nord - D5
C/Gran Capità s/n, 80034-Barcelona,
SPAIN

{magali.maziere|francoise.chassaing}@francetelecom.fr
{oster|philippe}@gps.tsc.upc.es

ABSTRACT

This paper examines the problem of segmentation and tracking of video objects for a content-based information retrieval context. Segmentation and tracking of video objects plays an important role in index creation and user request definition steps. The object is initially selected using a semi-automatic approach. For that purpose, a user-based selection is required to define roughly the object to track. In this paper, we propose two different methods in order to allow an accurate contour definition from the user selection. The first one is based on an active contour model which progressively refines the selection by fitting the natural edges of the object while the second one used a binary partition tree with a "marker and propagation" approach. The video object is thus tracked by using a hybrid structure alternately combining a hierarchical mesh for the motion estimation between two frames and a multi-resolution active contour model. This contour model is derived directly from the mesh boundaries in order to reposition the snake's nodes onto the natural edges of the object. The object-based segmentation associated to the object tracking allows relevant descriptors to be built for a future matching purpose.

1. INTRODUCTION

Video-based services are rapidly emerging as the next challenge for Internet-based applications. Since the networks carry more and more multimedia information, it becomes crucial to facilitate the access and the navigation for such video data. Some powerful content-based information retrieval systems may solve the problem of data superabundance. Such a system can be split into three parts: index generation, request translation and request/index matching. The index generation is an off-line process, it captures some content-based features (also called descriptors) linked to the video objects. The request translation consists of making the user selection comprehensible by the matching system. The objective of the matching part is to compute similarity distances between the query descriptors and the content-based index descriptors associated with the database. A series of images or video sequences are given as the result in decreasing order of similarity between the request and the index. In this paper, our main objective targets both the index creation and the user request definition step, which leads to the segmentation and tracking of video objects. The user selection is semi-automatic in the sense that the user roughly defines the object with a standard input device like a mouse or a graphic tablet. Then an object-based segmentation algorithm allows the video object to be separated from the background using this selection sketch. This segmentation should define accurately the contours of object to track, and it is used for both initialization of request and index descriptors generation. The object tracking algorithm, also applied to both request and index descriptors generation, is used to extract some relevant descriptors for a future matching purpose. Thus, the object tracking should capture the entire temporal contribution for the object of interest. This is why it is necessary to track the selected object for the longest time possible. The index is mainly built off-line by an operator, but the user may decide to manually incorporate a given object into the index database, so that the index may be completed. The present tracking method can thus be applied for an index construction purpose or for an accurate user request construction.

This paper is organized as follows: section 2 gives some feasible applications using object-based segmentation and tracking algorithms. Section 3 presents the interactive object selections while section 4 explains our object tracking methods. Finally, some conclusions and perspectives are given in section 5.

2. FEASIBLE APPLICATIONS

As we can today "surf" on the networks via text-based hyperlinks, it becomes necessary to develop the same concept for the video media via object-based hyperlinks. So we can imagine a navigation system where different retrieval approaches can be used in order to navigate a set of videos: keywords, image-based links, audio and video content-based requests. The user may use audio and video information in order to improve his request. The audio request can be acquired from a microphone or generated from user-defined words. A text-to-speech synthesis is then used in or-

0-7803-6536-4/00/$10.00 (c) 2000 IEEE

der to translate the sentence and the similarity distance between this translation and the indexes are processed. The results of the search are presented as a set of hyperlinks, the user being able to play one scene or all the scenes one after the other. An interesting feature of such a navigation concerns the object-based hyperlinks. A user may be able to select with a standard input device an object of the video he is looking at. If this object is linked to other objects of the video database, the user may browse the associated videos. It may also be possible to view some complementary information attached to the selected video object. At present, most of the video production is manually indexed. With proposed technique, a semi-automatic process becomes possible. The operator should be able to control the tracking process and attach some additional descriptors and links if necessary.

3. INTERACTIVE OBJECT SELECTION

Image segmentation remains an open problem, whereas it is still an important research area in computer vision. Many techniques are available, and a good review of these techniques has been done by Nikhil et al [1]. Full automatic methods don't give great satisfaction because they depend on image contents and on the particular application. This is why a semi-automatic object-based segmentation method is chosen to select an object of interest. Two different methods are proposed in this paper: the first one is based on an active contour model while the second one uses a binary partition tree.

3.1. Active contour models

Even when the user contribution is slight, the method mainly leans on it at the beginning to separate the object from the background. The user roughly defines the external object contours with a standard input device like a mouse or a graphic tablet. This selection sketch is then iteratively refined using active contour models in order to accurately fit the natural edges of the object.

The snake models were introduced by Kass et al [2] and many methods have been developed up to now: active contour models, balloon models [3], or geodesic models [4]. Our method defines and minimizes an energy function given by equation 1. This function is split into two parts: $E_{internal}$ corresponds to the internal snake energy and permits the model contraction and $E_{external}$ refers to the image and is often tied to its edges.

$$E_{snake} = E_{internal} + E_{external} \quad (1)$$

3.2. Binary Partition Tree

An alternative to the active contour models may be the "marker and propagation" approach described by [5] and figure 2 shows corresponding application window. For that purpose, a Binary Partition Tree is used. This tree is a structured representation of regions that may be obtained from an initial partition. A Binary Partition Tree is created by keeping track of regions that are merged by a region based segmentation [6]. Starting from an initial partition, the algorithm merges neighboring regions following a homogeneity criterion until one single region is obtained.

The "marker and propagation" approach consists first of marking (as one can do with a fluorescent pen to highlight a few important words) the interior of the regions to be segmented and second, of performing a propagation from these markers to define accurately the contours of the object using a bottom-up approach. Propagation processes based on similarity between neighboring regions can be easily implemented in the Binary Partition Tree. Consider the example of figure 1. At the top of this figure, one can see that the user has manually defined two objects of interest (dark and grey). Corresponding nodes have been marked using the same label on the Binary Partition Tree. Propagation is thus performed using the fact that the most similar region of a node is its brother. Therefore, the brother's nodes of a marker are marked with the same label. Result of the merging process is represented by marking their father. Of course, this propagation can only be done when the brother is not in conflict with the marker, that is if none of the brother's descendants have been assigned to a different marker. On bottom of figure 1, the result of the propagation on the tree and at the associated object segmentation is shown.

Once the object has been segmented, a contour extraction algorithm is used in order to obtain the external boundary. Then the contour is regularly sampled using a node placed at each of these samples. This representation is later used to perform the tracking.

4. OBJECT TRACKING

The video object tracking process captures the entire temporal contributions of an object. This step allows some relevant descriptors to be built for future matching. It is therefore significant to track the video object as long as possible, while being tolerant with its geometric deformation. We assume that the object color distribution remains constant during the tracking, and that small and progressive deformations and/or motions are applied to it.

Recent works have combined snake models and motion estimation. Paragios et al [7] used geodesic active regions with a constrained optical flow while Pasqual et al [8] used snake models with flow density. Our method uses a hybrid structure combining an active contour model and a hierarchical mesh well suited for the motion estimation process (see [9] [10] [11]). Such a mesh structure is associated with a hierarchical motion estimation approach: the coarse mesh level evaluates global motion while the finest levels estimate the internal motions of the video object.

4.1. Mesh hierarchical creation

An initial mesh is built from the nodes of the contour model and a regular spatial sampling on its interior region. For that purpose, a Delaunay triangulation process constrained to the region boundaries is used. The next mesh level is built according to a node-based sub-sampling and an edge-constrained Delaunay triangulation. This process is iteratively applied until a given number of hierarchical meshes is obtained. Figure 4 shows an example of this hierarchical with four meshes.

4.2. Motion estimation

The object motion is estimated using an affine model (equation 2). This affine model is able to represent most of the geometric deformations that an object may have. In the framework of triangular meshes, estimating this model is equivalent to computing a motion for each node of the mesh. The motion is initially estimated on the coarsest mesh resulting in the motion vector of nodes $p(x,y)$. The motion $\vec{d}$ inside each triangle is directly interpolated using barycenter coordinates Ψ of its three referring vertices i, j and k (equation 3). A global differential method thus propagates this motion estimation on the finer levels. It is based on a Gauss-Newton optimization algorithm regularized by Levengerg-

Marquartd method (see [9]) and minimizes the displaced frame difference between two successive frames (equation 4).

$$\begin{cases} u = ax + by + c \\ v = dx + ey + f \end{cases} \quad (2)$$

$$\vec{d}(p) = \sum_{n=i,j,k} \Psi_n^e(p).\vec{d}_n \quad (3)$$

$$DFD(x,y) = I_t(x,y) - I_{t+1}(x+u, y+v) \quad (4)$$

This motion estimation is applied to the mesh nodes in order to obtain a first approximation of the object at the current frame. An active contour model is then built from the finest level of the mesh hierarchy, in order to improve the object contour and thus the object tracking.

4.3. The active contour model

The active contour model is generated from the edge nodes of the boundary of the finest mesh level. The model evolution thus allows the edges of the object to be improved.

The object tracking restarts from the current contour model as described by section 4.1. Figure 3 shows different steps of the global object tracking process.

5. CONCLUSION

In this paper, we have discussed the interest of using segmentation and tracking of video objects algorithms for a content-based information retrieval context. The two important steps concern the interactive object selection and the object tracking. The user initializes the process with a rough selection of the object. An algorithm is then applied in order to obtain an accurate contour of the object. The video object is tracked as long as possible by a hybrid method combining a hierarchical mesh for the motion estimation between two successive frames and an active contour model in order to improve the edge fitting.

The object based tracking method gives good results for particular objects and allows us to define the object internal motion. It is sometimes important to characterize internal motion in order to control and detect some anomalies of the object. This method needs large objects without a contraction zone as for example a man holding a fishing rod. The problem comes from the subsample method. It assumes the mesh is fine enough according to the mask. If the object of interest is too small, the boundary of the meshes is not near enough to the object edges (see figure 4), but thanks to the snake method, we can accurately fit the natural edges of the object.

For the global indexing project, we need an object independent tracking method. So we decided to combine motion estimation with the affine model and snake method. Results show robust object-based tracking during several images. For example on the dancer video, this method tracks the object defined by the red and white dancer on two hundred first image. Moreover, execution speed is one second per image on a PII350.

Our future work concerns the natural next step of presented work: the building of relevant color and texture descriptors in order to characterize an object and the whole image, also called support. We would like to try to capture the relevant information of support into a set of descriptors which memory size is highly smaller than the support.

6. REFERENCES

[1] N.R Pal and S.K Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1295, 1993.

[2] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes : active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[3] Laurent D. Cohen. On active contour models and ballons. *Computer Vision Graphics and Image Processing : image understanding*, 53(2):211–218, 1991.

[4] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[5] Philippe Salembier and Luis Garrido. Binary partition trees as an efficient representation for filtering, segmentation and information retrieval. *ICIP, Chicago*, october 1998.

[6] Luis Garrido and Philippe Salembier. Extensive operators in partition lattices for image sequence analysis. *EURASIP Signal Processing*, pages 157–180, april 1998.

[7] Nikos Paragios and Rachid Deriche. Geodesic active regions for motion estimation and tracking. *Internationnal Conference in Computer Vision*, 1999.

[8] Ajith Pasqual and Kiuoharu Aizawa. Tracking and shape extraction of a moving object in video sequences using feature integration and snakes. *Very Low Video Bitrate*, pages 113–116, 1999.

[9] Patrick Lechat. *Représentation et codage de séquences vidéo par maillages 2D déformables.* PhD thesis, University of Rennes, october 1999.

[10] Patrick Lechat, Michael Ropert, and Henri Sanson. Hierarchical mesh-based motion estimation using a differential aproch and application to video coding. *EUSIPCO*, 4:2081–2084, september 1998.

[11] Patrick Lechat, Nathalie Laurent, and Henri Sanson. Suivi d'objets vidéo par maillage hiérarchique. *GRETSI*, septembre 1999.

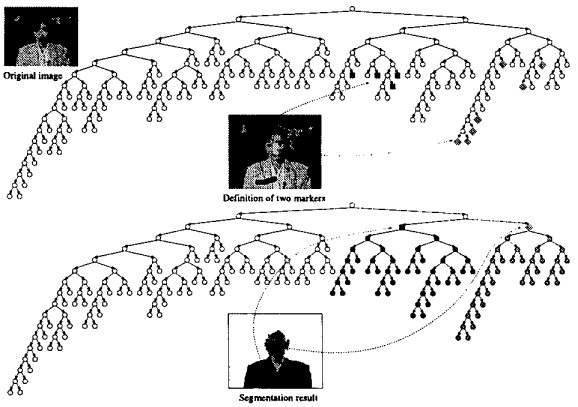

figure 1: Example of "marker and propagation" approach: above, two markers (dark and grey) are defined and they are associated to the tree nodes. Below, the result of the propagation process on the tree and the associated segmentation is represented.

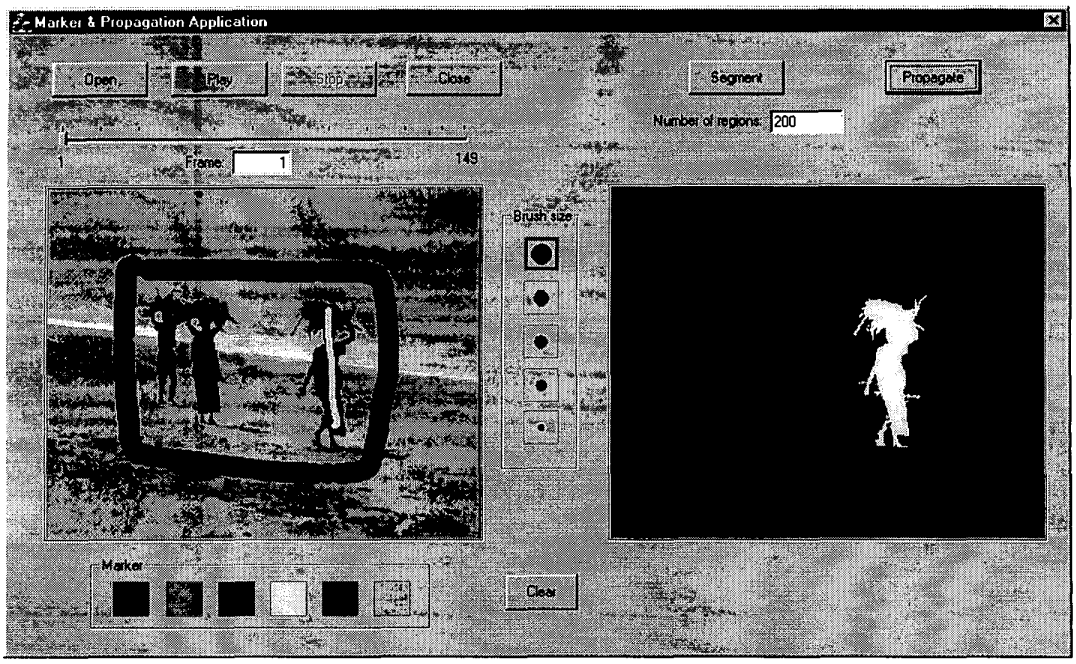

figure 2: This application segments an image according to the desired number of regions, builds automatically the Binary Partition Tree and allows us to mark the regions of interest. On the left window, the application displays the segmentation image and one may select a marker region with color pen. The size of this pen brush can be defined by the user. The propagation result is displayed on right windows. It is not the case here but some regions may appear in black if they are not assigned to any marker. This happens when a conflict occurs during the propagation of the markers through the Binary Partition Tree.

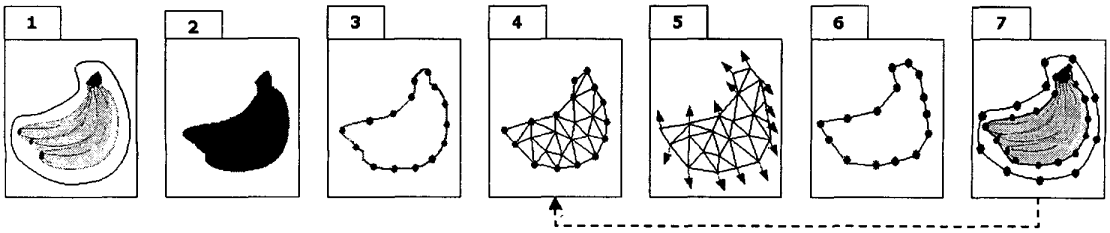

figure 3: Different steps of object tracking algorithm: 1.Selection sketch, 2.Initial segmentation, 3.Edges detection and placement of nodes, 4.Mesh generation, 5.Motion estimation, 6.Motion compensation and mesh to snake conversion, 7.Contour fitting with active contour models.

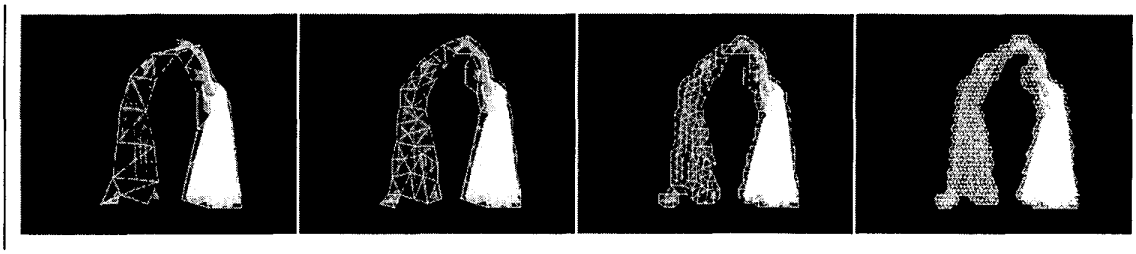

figure 4: Example of mesh hierarchical generation with four meshes. The active contour model is only generated from the edge nodes of the boundary of the finest mesh level ie the right one.