

Semantic Video Object Extraction Using Four-Band Watershed and Partition Lattice Operators

Daniel Gatica-Perez, *Student Member, IEEE*, Chuang Gu, *Member, IEEE*, and Ming-Ting Sun, *Fellow, IEEE*

Abstract—We conceive the problem of multiple semantic video object (SVO) extraction as an issue of designing extensive operators on a complete lattice of partitions. As a result, we propose a framework based on spatial partition generation and application of optimal operators on the generated partitions. Based on a statistical analysis of the watershed algorithm, we develop a multivalued morphological spatial segmentation method that incorporates an edge-driven marker extraction algorithm and a growing method which integrates both color and edge information. Having embedded the problem in the partition lattice framework, we propose a spatio-temporal regional maximum likelihood operator for extraction purposes. Some theoretical properties of the operator are established. Experimental results on several MPEG-4 test video sequences show that our scheme improves the precision of the extracted SVO boundaries compared to traditional watershed algorithms and provides accurate tracking of multiple SVOs in both static and moving camera scenarios. Furthermore, this scheme can be extended to deal with more general interactive video authoring systems.

Index Terms—Mathematical morphology, multivalued watershed, partition lattice operators, semantic video object.

I. INTRODUCTION

THE extraction of meaningful entities—objects in the real world—from visual data is one of the highest aspirations in image analysis. With the development of content-based multimedia systems, the need for automatic or semiautomatic generation of semantic objects from natural video sequences is beyond any doubt. Three applications are the main immediate beneficiaries of the development of video object extraction algorithms: 1) selective compression for storage and transmission; 2) personalized multimedia applications that allow for search, manipulation, and composition of object-based, hybrid video content; and 3) visual communication services that bring people together to interact in either working or entertainment environments. Such paradigms are properly specified in MPEG-4 and MPEG-7, the next generation of international multimedia standards.

SVO extraction can be considered as a process of segmenting and tracking *arbitrary* collections of image regions—scene objects—with *pixel-wise accuracy*. SVOs are essentially *human*

abstractions that are not invariant either in spatial features or in motion; given this set of challenging requirements—loose in the definition of what an object is, ambitious with respect to the final result quality—it is not surprising that SVO extraction is a difficult task to accomplish. In particular, pixel-wise accuracy is necessary for several MPEG-4 applications and, although not mandatory for MPEG-7 video indexing applications, it can eventually improve their robustness and performance.

Recently, a large number of approaches, working under different perspectives and combining old and new techniques in feature extraction, segmentation and tracking, have been proposed. For *automatic* extraction of moving video objects, the MPEG group merged techniques based on morphological spatio-temporal segmentation and higher-order-statistic motion detection [7]. Binary edge moving object models were defined in [18] and tracked using a modified version of the Hausdorff distance [23]. However, people have now agreed on the need for *semiautomatic* methods [8], [13], [15], which are based on user-assisted definition of SVOs followed by object tracking. Most of the firmly established trends in image analysis have been explored for this purpose [8], [13], [15], [17], [31]. Creation and tracking of 2-D meshes of arbitrary shape was the direction developed in [31]. Active contours constitute a solid research area [3] that has been integrated in many of the existing approaches [6], [31]. Corners and lines are commonly used as rigid object representations [35] but are limited when dealing with deformable objects. Different sorts of object templates have been widely employed for object tracking [32]. Finally, representations based on 2-D regions have been used in the past for tracking [19], and more recently for SVO extraction [13], [17].

A survey of the previous work in this area points toward a generic four-step SVO extraction model, that comprises many of the existing methods: (1) SVO structure definition; (2) SVO computation; (3) SVO tracking, and (4) SVO postprocessing. In the first step a *SVO representation scheme*, the image features that define the object, is selected. Furthermore, if interactivity is allowed, a set of user-selected features is also provided, usually in the form of one or more contours, blobs, or context information. Binary models, meshes, corners, templates, object contours, and 2-D regions are all examples of these representations. The second and third steps of the generic model are highly related: the iterative computation and tracking of the SVO structure generates the video objects at each frame of the sequence. A large number of techniques are available, depending on the selected representation. Finally, the fourth step is usually necessary to deal with the inaccuracies produced at any of the previous stages.

Manuscript received September 3, 1999; revised October 6, 2000. The work of D. Gatica-Perez was supported by the Fulbright-CONACYT Ph.D. scholarship program, and by the National University of Mexico.

D. Gatica-Perez is with the Department of Electrical Engineering and the Human Interface Technology Laboratory, University of Washington, Seattle, WA 98195 USA.

C. Gu is with Microsoft Corporation, Redmond, WA 98052 USA.

M.-T. Sun is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: sun@ee.washington.edu).

Publisher Item Identifier S 1051-8215(01)03822-8.

When 2-D regions are selected as the SVO structure, two factors affect the quality of the final extraction result: the precision of the spatial partition and the selected tracking technique. On one hand, a good spatial segmentation technique should preserve all the contours that define the true objects, as human perception is sensitive to artifacts in object borders. On the other hand, the tracking process is responsible for discerning between past and present information in order to keep an accurate SVO representation along time. Considerable research on region-based object tracking has been done [9], [11], [19], [34]. Most of it can be generically labeled as *partition prediction-adjustment* methods, in which the problem can be stated as follows: given the set of SVOs, the image sequence, and the last available partition, how does this partition need to be updated/deformed to generate the current one? The process of partition updating, however, is not an easy task if pixel-wise accuracy is required, as discussed elsewhere [13], [17]. Noisy motion information is usually the problem in this process, which often introduces inaccuracy and uncertainty in defining the number and boundaries of the regions that compose the video objects. In addition, some of these adjustment procedures unfortunately rely on heuristics, and/or cannot handle the real-life case of multiple objects.

Based on a 2-D region representation, and building on the generic four-step SVO extraction model, this paper proposes a framework for the second and third steps, defining it as an issue of *partition lattice operators* [27], and characterized by *spatial partition generation* and the application of one or more *spatiotemporal partition operators*, respectively. The formalism of operators on lattices of partitions has been previously used in digital video analysis [10], although not specifically for SVO extraction.

For the partition generation phase, based on the results of statistical analysis of the contour localization properties of the watershed algorithm, we have developed a new four-band multivalued morphological segmentation method that improves the performance of the traditional techniques by introducing an edge-based marker extraction step, and by integrating edge and color information in a multivalued model. As a result, it provides better precision when used for SVO extraction.

In our framework, the extraction process is then accomplished by designing extensive operators on the generated partitions. We present a new optimal operator based on spatio-temporal regional maximum likelihood. Some of its theoretical properties are established. When the proposed approach is integrated in a semiautomatic system, it is able to accurately extract multiple, arbitrary SVOs in still and moving camera video clips. Both objective and subjective evaluation on several test image sequences illustrate its performance.

The rest of the paper is organized as follows. Section II presents a general overview of our methodology, and discusses previous related work. Section III first discusses the statistical analysis of the watershed algorithm, and then describes our partition generation method. Section IV illustrates the SVO extraction approach by introducing the partition lattice operator framework. Section V shows results for several MPEG-4 test sequences. At the end, Section VI provides some concluding remarks.

II. PROPOSED APPROACH

Our goal is to develop a systematic approach for SVO extraction that:

- 1) represents a theoretical framework in which different operations can be defined, and their properties can be analyzed;
- 2) allows for real-life applications, so that multiple, deformable, static, or moving SVOs can be extracted with pixel-wise precision.

As stated in the previous section, a region-based SVO representation was adopted. Therefore, a complete lattice of partitions constitutes an appropriate framework to address SVO extraction [27], [29].

The elements of a partition lattice Π are all the possible partitions of the support \mathcal{E} of an image, ranging from the *finest* partition (in which every pixel is a region) to the *coarsest* one (in which the whole image constitutes a single region). The existence of an *ordering relation* allows for comparison between those partitions in Π that build a hierarchy [10]. A lattice of partitions is a particular case of a complete lattice, the general algebraic structure on which mathematical morphology is developed [14], [27]. A complete lattice represents the proper benchmark to analyze theoretical properties of the operators that can be defined on it. The reader is referred to the Appendix, where we provide some basic definitions used in this paper, and to [27] and [14] for details.

The extraction of the SVOs of a scene corresponds to one special case of partition of the image support. This can be defined as follows.¹

Definition 1: Let $\mathbf{I} = \{\mathbf{I}^t \mid t \in \mathbb{Z}\}$ be a multivalued image sequence, with domain $\mathcal{E} = \mathcal{D}(\mathbf{I}^t) \subset \mathbb{Z}^2$. Let $P^t = \{R_i^t, i \in \{1, \dots, N^t\}\}$ denote a partition of \mathcal{E} at time t . The j th semantic video object of the scene depicted in \mathbf{I} (consisting of M objects) is defined by $SVO_j = \{SVO_j^t\}$, where

$$SVO_j^t = \bigcup_{i=1}^{N_j^t} R_i^t, \quad (1)$$

In MPEG-4 terminology, SVO_j^t represents the j th video object plane (VOP) at time t , each composed of N_j^t regions of P^t ($\sum_j N_j^t = N^t$). Equation (1) naturally allows for the definition of multiple SVOs. The associated partition of SVOs at time t , denoted by P_{SVO}^t , is the collection

$$P_{SVO}^t = \{SVO_j^t, j \in \{1, \dots, M\}\}. \quad (2)$$

We based our methodology on two premises.

- 1) SVO extraction can be achieved by generating accurate spatial partitions P^t that do not depend on noisy motion information and therefore preserve the true object contours, and by finding optimal operators on a lattice of partitions, $\psi^t : \Pi \rightarrow \Pi$, such that each region $R_i^t \in P^t$ is mapped to the corresponding SVO in the scene. The notation ψ^t indicates that, in general, the operators can be time-adaptive.

¹Multivalued images and random variables are both denoted by bold letters. The meaning should be clear from the context.

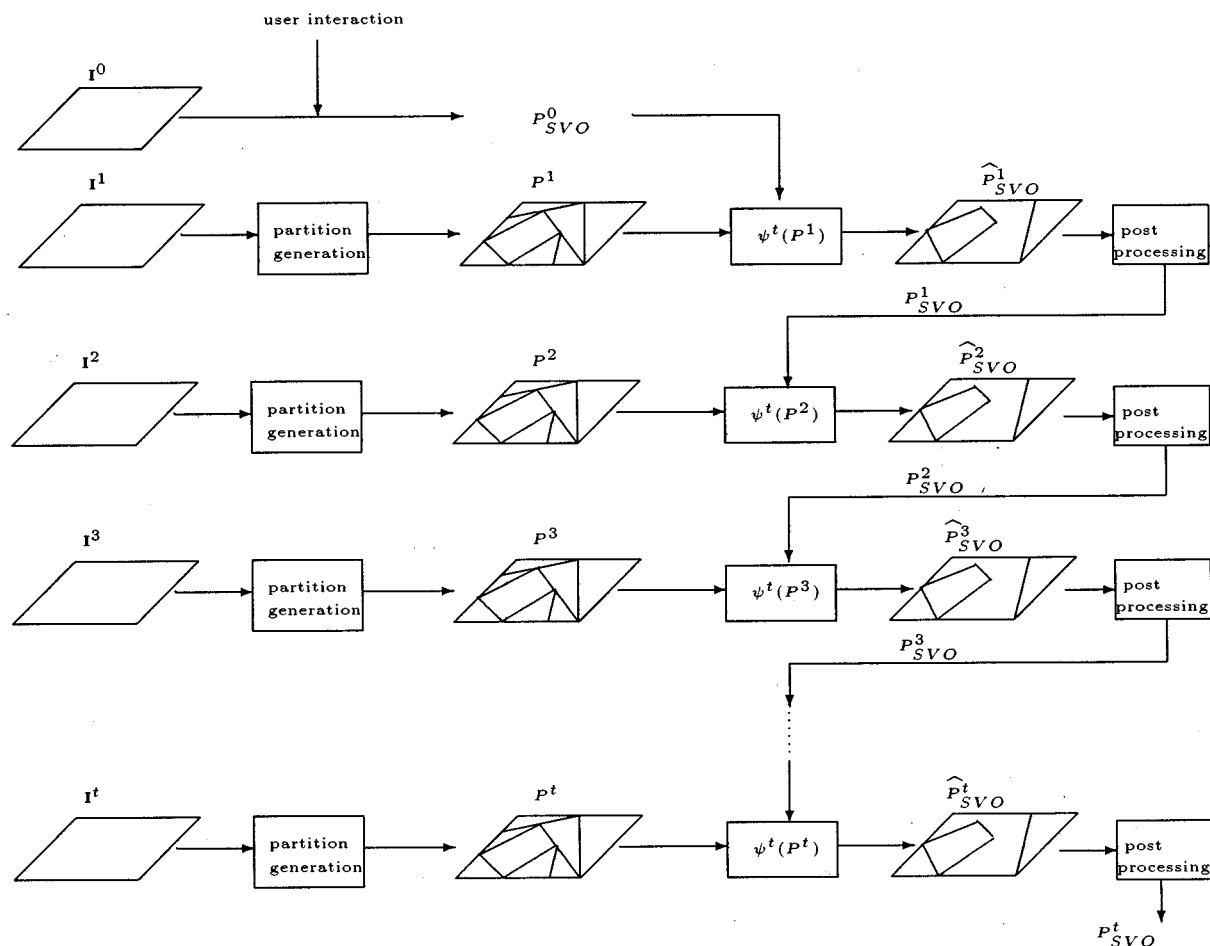


Fig. 1. Framework for SVO extraction, for the case of minimum user intervention, as inscribed in the four-step generic model. An accurate partition generated by multivalued morphological segmentation followed by the application of optimal extensive partition lattice operators create the partition of SVOs at each frame.

- 2) Interactive introduction of semantics is essential. The user should be allowed to implement one or more of the following functions: 1) definition of the original SVOs; 2) definition of new objects; 3) creation of a multi-view representation; 4) correction of the automatic results; and 5) specification of context. Increasing work in some of these directions is currently being done [4], [8], [13].

SVO extraction is then conceived as a *combined process of spatial partition generation and subsequent application of partition lattice operators, with user intervention at (possibly) different instants.*

In our approach, the partition generation step is responsible for providing an accurate division of the image into contour-preserving regions R_i^t , such that the frontiers between objects can be discerned even if they are of similar color. Moreover, once a partition P^t is generated at each time t , the problem becomes the *construction of P^t_{SVO} from P^t by specifying spatio-temporal statistical criteria, and by introducing temporal information that allows the implementation of the tracking function.* This information is represented by a *temporal reference SVO partition set*, denoted by TR , composed of the SVO partitions of the scene at different time instants, i.e., partitions that correspond to different scene views. The partition reference set is then expressed as $TR = \{P^t_{SVO}, k \in \{1, \dots, K\}\}$. For example, if $K = 1$, and $t_K = t - 1$, then $TR = \{P^t_{SVO}\}$, which means that the

generation of the current SVO partition will depend on information provided by the previous one. If $K > 1$, the decision for the construction of the current SVO partition will include information from multiple instances. Note that the partitions in TR can be computed either by off-line user interaction or automatically as part of the extraction process. We now define a SVO extraction operator.

Definition 2: Let Π be a complete lattice of partitions of $\mathcal{E} = \mathcal{D}(I^t)$. Let $P^t \in \Pi$, and $TR = \{P^t_{SVO}, \dots, P^t_{SVO}\}$. An SVO extractor operator is a mapping $\psi^t_{TR} : \Pi \rightarrow \Pi$, so that

$$P^t_{SVO} = \psi^t_{TR}(P^t). \quad (3)$$

The explicit dependence on TR will be usually omitted as this set is used as a reference. With this formulation, several tracking schemes can be formulated: single-view ($K = 1$) or multiview ($K > 1$); causal ($t_k < t$) or noncausal ($t_k > t$). Theoretical properties of the operators can be obtained in the complete lattice framework.

Fig. 1 illustrates the case of minimum user interaction in our approach, in which we limit the user to specify only the initial SVOs (the theory and results obtained for other types of user interaction will be reported elsewhere). We inscribe our scheme in the generic SVO model presented in Section I as follows.

- 1) *SVO definition*: a user-defined SVO initial partition P_{SVO}^0 is generated from the first frame of the sequence.
- 2) *SVO computation*: each frame is partitioned using our proposed four-band (color+intensity edges) morphological multivalued spatial segmentation method. The corresponding P^t is generated.
- 3) *SVO tracking*: each partition is processed with a new spatio-temporal regional maximum likelihood extensive operator that generates the current partition of SVOs, defining $TR = \{P_{SVO}^{t-1} : \hat{P}_{SVO}^t = \psi_{TR}^t(P^t)\}$.
- 4) *SVO postprocessing*: small details are refined and smoothed. The final partition of SVOs is generated: $\hat{P}_{SVO}^t \rightarrow P_{SVO}^t$.

Before describing our work in detail, we will elaborate on its connections to previous work in the next subsection.

A. Relation To Previous Work

The concept of lattices of partitions was described in [29], and integrated in the complete lattice approach of mathematical morphology in [27], where some partition operators for segmentation were proposed. The formalism was implicitly used for segmentation purposes in [25], that proposed the creation of a hierarchy of partitions (which correspond to extensive operators in a lattice of partitions) for coding purposes. Recently, the formalism was explicitly employed in image sequence analysis to highlight algorithmic similarities and differences between morphological connected operators and extensive operators in lattices of partitions, and to propose region-merging algorithms for still-image segmentation and filtering [10]. In those operators, regions were merged with a one-region-at-a-time iterative methodology based on three elements: a region model, a merging order, and a merging criterion. The region models allowed for different schemes, including gray-level and motion-oriented segmentation. Additionally, a similar methodology was applied to image sequences, to perform motion segmentation based on forward region tracking. Such a scheme relied on a three-level hierarchy, with a fine, flat-zone partition in the bottom, a gray-level partition (that iteratively merged some regions from the flat zone partition) in the middle, and a motion partition on the top. A sophisticated procedure was used for forward region tracking. We consider that is the work that bears the most similarity to our work, although there exist several fundamental differences between them.

- 1) Motion segmentation and SVO extraction are not equivalent, because SVOs are usually not motion-homogeneous. Our proposed operators do not require that the objects are motion-consistent.
- 2) The work in [10] does not explicitly deal with SVO extraction, although the region-tracking technique can be useful for that purpose. A more recent work addresses the problem of SVO extraction [17], but both the spatial partition generation and tracking techniques are different from ours.
- 3) The region-merging operators of [10] do not specify what type of merging model/order/criteria can be used so that real contours are preserved. Our spatial generation pro-

cedure only relies on region merging as a postprocessing stage of the spatial generation algorithm.

- 4) Unlike the operators in [10], which are iterative, and rely on a strong connectivity property, our operators classify each region independently, and are not iterative.

The fundamental coincidence consists in the use of the algebraic formalism of partition lattice to represent and analyze the dynamic content of video sequences, and to study the theoretical properties that can be derived from it.

III. PARTITION GENERATION: FOUR-BAND MULTIVALUED WATERSHED

The family of morphological methods for segmentation are generically referred to as watersheds [1]. Methods that operate both on gradient images and on grey-level/color information have been proposed. In particular, a region-growing algorithm based on the watershed transformation consists of three steps: *image simplification*, *marker extraction*, and *decision* [25]. The first step facilitates the segmentation process as it eliminates some noise and small details while preserving strong image contours. The second step, marker extraction, generates a set of connected pixels from the simplified image. The set of markers represent the initial regions from which the segmentation is built in the third step, decision, which is the watershed transformation itself. In this stage, region boundaries are decided by a simulated immersion process that floods the whole image from the extracted markers.

Marker extraction and decision determine the number of regions and the precision in boundary localization of the final segmentation, respectively, and thus are critical procedures. In particular, the problem of determining the number and position of the markers has been addressed by several researchers. In [1] and [33], markers are extracted by selecting gray-level *h*-domes. Another method that has gained popularity is the *flat-zone approach*: all connected components in the simplified image of size larger than the size of the structuring element used for morphological simplification are considered as valid markers [25]. It has proven to be useful for segmentation-based image coding, but presents limitations when used for SVO extraction, in which we need to preserve all image contours -weak or strong- for precise object localization. Finally, off-line marker extraction can be useful but it is restricted to relatively simple imagery. In spite of the previous efforts, an analysis of optimal marker extraction remains elusive, although authors agree on the fact that this process should be intelligently performed [20].

We believe that a statistical analysis of the border localization properties of the watershed algorithm should provide some clues about the criteria for good marker selection, with SVO extraction as the ultimate goal. In fact, the authors do not know of any similar studies previously reported in the literature. In the following section, we present such an analysis for a particular watershed algorithm that works directly on the gray-level signal [25] for a ramp edge+noise model, to evaluate the objective quality of the obtained partitions. Our study suggests that, for weak edges, the optimum position for the markers occur when they are symmetrically located on both sides of the edges and arbitrarily close to them.

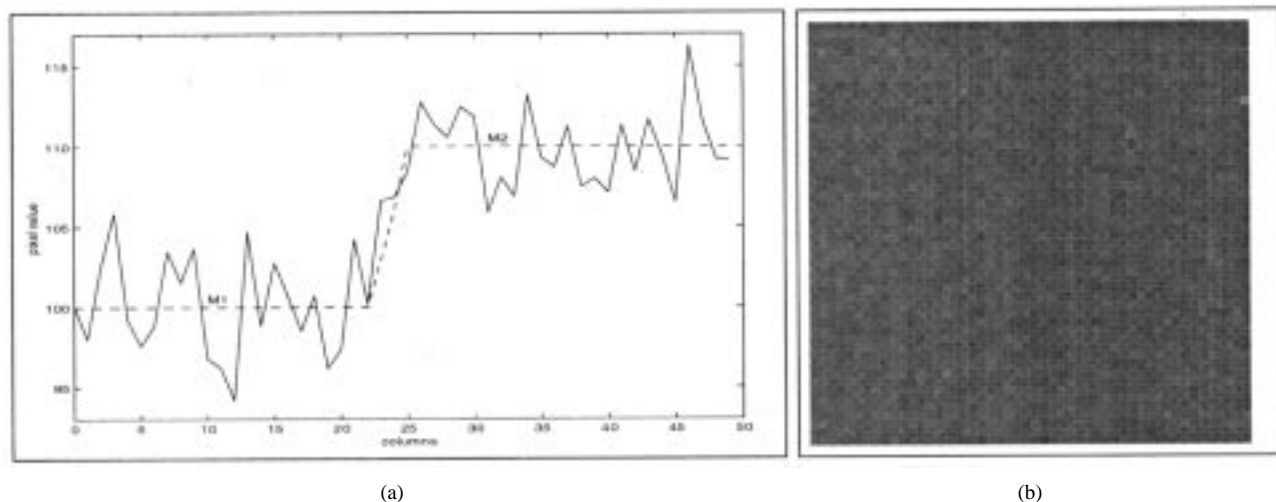


Fig. 2. Ramp edge model. (a) Profile of a 50×50 two-region image with a 3-pixel width and weak vertical edge. The two regions have mean values 100 and 110 (dashed line), and are immersed in white noise (continuous line). The markers M_1 and M_2 are randomly positioned on both sides of the edge. (b) Synthetic image with the proposed ramp model. An ideal segmentation algorithm would produce a vertical line dividing the two regions exactly at column 23.

A. Statistical Analysis of the Watershed Algorithm

The watershed algorithm can be efficiently implemented using a hierarchical queue, in which the element that is extracted at each time is the one that arrives *the earliest* to the queue of *the highest* priority [1]. This double-ordering scheme is the key concept for the process, that consists of two steps: initialization and flooding. First, the queue is initialized by inserting the marker pixels in the highest priority queue, and all markers are labeled as initial regions. Then, until the whole queue is empty, a pixel x is extracted from the queue and assigned to the most similar neighboring region, using a distance criterion $d(x, R_i)$ as the classification criterion. Its neighboring pixels $\{x_N\}$ are inserted in the queue with priority $d(x, x_N)$. Therefore, pixels that are more homogeneous to the existing regions are classified earlier.

To evaluate the border localization properties, we introduce an image model with controlled edges. Furthermore, we define a variable that reflects the accuracy in the segmentation, so that its statistical behavior can be obtained [30], and we can make inferences about marker extraction.

For the image model, let $I: \mathcal{E} \rightarrow \mathcal{R}$ be a gray-level image with a vertical ramp edge of slope m between two regions represented by two randomly positioned markers M_1, M_2 of unitary width. The image is immersed in Gaussian white noise n of variance σ^2 . A profile of the model is shown in Fig. 2.

An ideal segmentation would divide the two regions exactly on the edge. In practice, the position of the watershed line is not only a function of the initial position of the markers, the noise amplitude and the edge amplitude, represented by its slope, but also of the order in which pixels are treated during execution, because the watershed transform is not a local concept [22]. For initialization, we introduced the marker pixels in the queue in the usual raster order. This has an effect on the segmentation result, that is small in natural images, but might be evident with images of reduced dynamic range [22].

Let X be the set of all connected subsets in \mathcal{E} and let $x_1, x_2 \in X$. We define a gray-level distance criterion as $d(x_1, x_2) =$

$[\lceil \bar{x}_1 - \bar{x}_2 \rceil]$, where $\bar{x} = \sum_{i \in x} I(i) / \text{card}(x)$, and $\lceil \cdot \rceil$ represents truncation, so that the queue deals with integer priorities. Let w denote a random variable that represents the final position of the watershed line. Due to the complexity of the hierarchical queue and the dependency of initialization, the conditional pdf $p_w(w | n, M_1, M_2, m)$ cannot be obtained in closed form, but approximated. A simplification in the conditional pdf can then be done by introducing a *edge-to-noise strength ratio* term, defined by $f = |m|/\sigma$. Note that the combination of noise and a small edge amplitude increases the uncertainty in the watershed process and produces segmentation errors. The conditional pdf $p_w(w | f, M_1, M_2)$ can provide an evaluation of the quality of the obtained partitions.

Three experiments were designed to provide clues for intelligent marker extraction.

- 1) *Dependency on the markers position.* Fig. 3(a) and (c) show six 50×50 images in which the markers (shown in white) have been asymmetrically placed with respect to a vertical *weak* edge ($f = 0.7$). The markers M_2 have been positioned close to it, while M_1 change position. Fig. 3(b) and (d) illustrate the segmentations obtained when the markers in Fig. 3(a) and (c) are used (the watershed line is shown in white). On one hand, a bias toward M_1 , which come earlier in the scanning direction, is observed. This is the effect of queue initialization. On the other hand, the farther M_1 are from the edge, the worse the edge localization. This effect also applies to real images: *weak intensity edges*, and *absence of markers* are responsible for many segmentation errors; such situation should be minimized in SVO extraction. The approximated conditional pdfs for M_2 fixed, and various values of M_1 are shown in Fig. 3(e). Note that the peak of the distributions shifts to a better position as the markers M_1 come closer to the edge.
- 2) *Dependency on the symmetry of the markers position.* Fig. 4(a) shows the estimated pdfs when M_1 and M_2 are placed at a decreasing symmetric distance from the edge $d_o = (M_2 - M_1)/2$. The symmetry is also evident in the

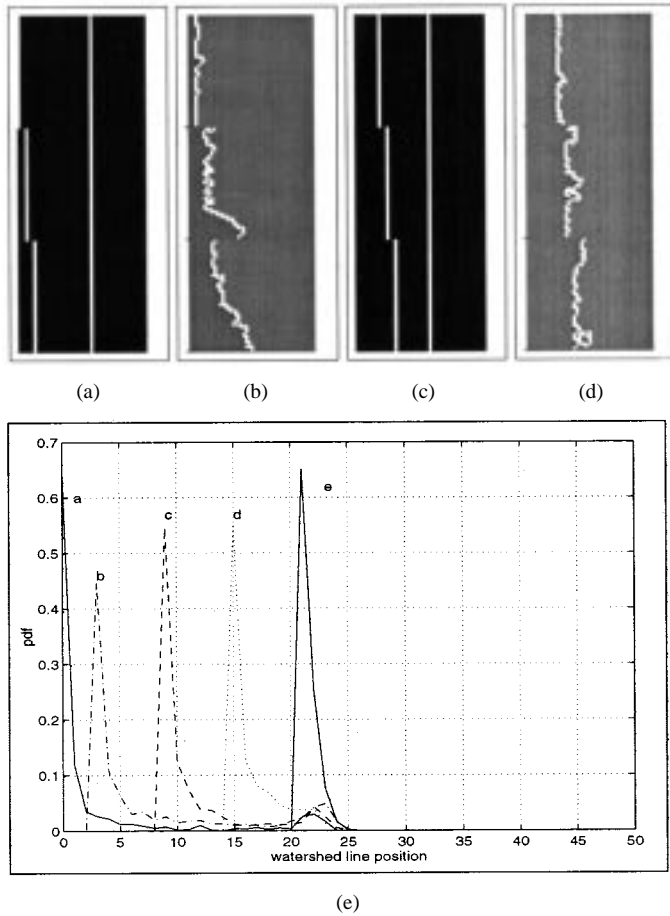


Fig. 3. Statistical analysis of the border localization properties of the watershed algorithm for a weak vertical ramp edge ($f = 0.7$, $w_{ideal} = 23$). (a, c) Six 50×50 synthetic marker images. The markers on the right of the edge have been positioned close to it ($M_2 = 24$), while the ones on the left (M_1) change position. (b, d) Watershed segmentations of the image in Fig. 2(b), when the markers in (a, c) are used. The farther M_1 from the edge, the worse the segmentation. (e) Conditional density functions $p(w|f, M_1, M_2)$ for $M_1 = 0, 2, 8, 14$ and 20 (curves a-e). As the markers M_1 come closer to the edge, the peak of the distributions shifts to w_{ideal} , improving the border localization.

conditional distributions. More accurate conditional expected values for w are obtained as d_o decreases. Table I illustrates these results.

- 3) *Dependency on f* . The distributions in Fig. 4(b) correspond to varying f , and fixed asymmetrical positions of the markers. As f increases, the peak of the pdf shifts from a value corresponding to an erroneous watershed line to the correct one (observe the trend a→d). For large values of f , the position of the markers is not crucial, as long as there are markers for each region.

These experiments, although not exhaustive, confirm an intuitive result: if the space where the growing process from the markers is restricted to small areas around the real edges, the segmentation results are of good quality. Therefore, even though the marker position is not crucially important for strong edges, constraining the propagation space by introducing markers close and centered around the true edges will improve the segmentation in those cases of similar color between different objects.

This result was integrated in the partition generation algorithm described in the following subsection.

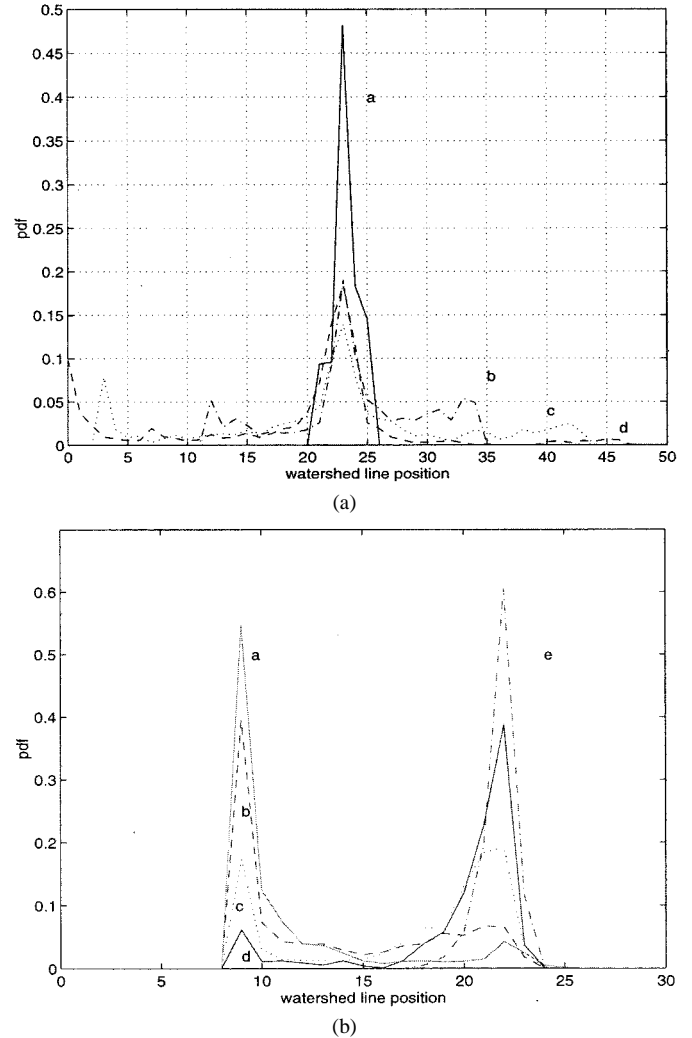


Fig. 4. (a) Conditional pdfs $p(w|f, M_1, M_2)$ for a weak edge model and markers *symmetrically* placed on both sides of the edge, at a distance $d_o = (w|f, M_1, M_2)/2$. Curve a corresponds to the smallest d_o , curve d to the largest one, and curves b-c to intermediate values (see Table I). The closer the markers are to the edge, the more concentrated the corresponding pdf is, and a better border localization is obtained, as can be seen from the first and second order statistics. (b) Conditional distributions for increasing value of f (0.7, 1.0, 1.2, 1.3 and 1.5 for curves a-e, respectively), and fixed *asymmetrical* markers ($M_1 = 8$ and $M_2 = 24$). As f increases, the peak of the distribution gradually shifts a wrong watershed line position (curve a) to the correct one (curve e). For large f , the conditional pdf does not change anymore, which reflects the fact that for strong edges, marker position is not crucial as long as there are markers for each region.

TABLE I
STATISTICS FOR WATERSHED LINE POSITION FOR SYMMETRIC DISTANCE
BETWEEN MARKERS ($w_{ideal} = 23$)

Curve	d_o (pixels)	$E_w(w f, M_1, M_2)$	$\sigma_{w f, M_1, M_2}$
d	23	18.10	9.96
c	20	22.27	10.17
b	11	23.94	5.86
a	2	23.19	1.56

B. Four-Band Multivalued Segmentation

The multivalued mathematical morphology framework provides a way of integrating information from different sources

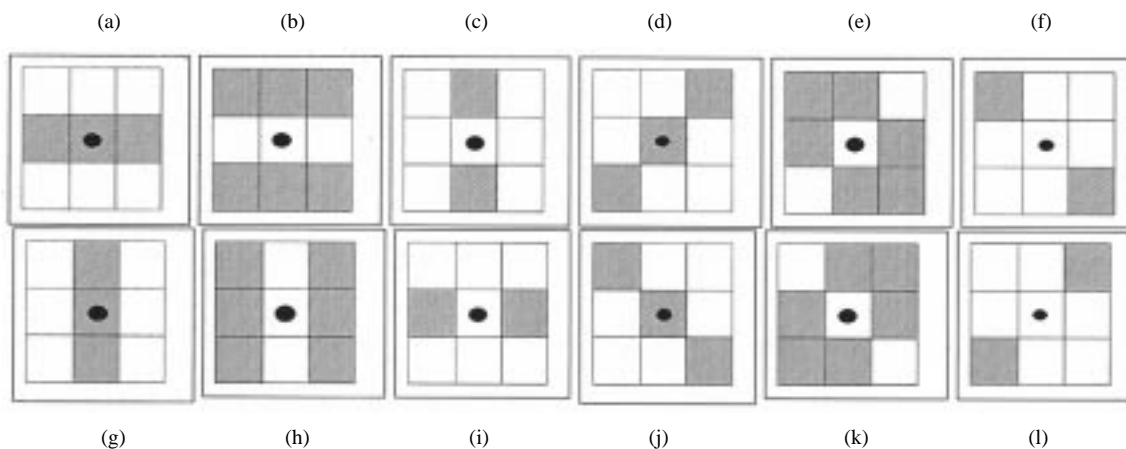


Fig. 5. Structuring elements for marker extraction (the black dot indicates their origin), for $N_{dir} = 4$ different orientations. (a-c) $\{B_1^j\}$ (0 rads). (d-f) $\{B_2^j\}$ ($\pi/4$). (g-i) $\{B_3^j\}$ ($\pi/2$). (k-l) $\{B_4^j\}$ ($3\pi/2$).

into one single image model [28], and has been recently used for image sequence processing and analysis [11]. Based on the results of the previous subsection, we propose a morphological partition generation method that combines color in the RGB representation and edge information in a four-band multivalued model

$$P^t = P^t(\mathbf{I}^t) = P^t(\mathbf{R}^t, \mathbf{G}^t, \mathbf{B}^t, \mathbf{E}^t) \quad (4)$$

where \mathbf{R}^t denotes the red band image at time t , and similarly for the other color bands, \mathbf{E}^t denotes a binary edge image, where 0 and 255 correspond to nonedge and edge pixel values, respectively. The explicit use of edge information improves the segmentation results in regions of similar color, which is a crucial point for finding precise boundaries between SVOs. The use of joint edge and color information is not new in image processing [16], [24], but it has not been used in morphological segmentation as we propose in this paper. Our method is based on four steps: simplification, edge-based marker extraction, decision, and postprocessing.

Simplification: Filtering-by-reconstruction with a small, 3×3 structuring element is applied [33]

Edge-Based Marker Extraction: Based on the previous analysis, we propose a nonlinear strategy that reduces the possibility of erroneous region growing during the watershed procedure due to absence of markers in the presence of weak intensity edges. Firstly, the edge image \mathbf{E}^t is obtained from the intensity image \mathbf{Y}^t . Edge detection has been a problem largely discussed in the literature. We decide to use the Canny edge detector [5] that is known to be accurate in edge localization, and is noise-robust. This technique produces a good number of thin, connected edge pixels, along with some other pixels associated to texture or large noise. We need to say that any other good algorithm that allows for the detection of weak edges can be used. Secondly, we apply binary morphological processing on \mathbf{E}^t to obtain \mathbf{M}^t , the image that consists of the markers extracted on both sides of detected edge pixels, along the normal to the corresponding edges. A reliable edge pixel is defined as one that has two or more neighboring edge pixels in 8-connectivity. The procedure consists of the detection of edge pixels by using the hit-or-miss

transformation followed by a dilation, for a set of possible edge directions N_{dir}

$$\mathbf{M}^t = \bigcup_{i=1}^{N_{dir}} (\mathbf{E}^t \otimes (B_i^1, B_i^2)) \oplus B_i^3 \quad (5)$$

where \otimes and \oplus denote hit-or-miss and dilation, respectively, and the sets of structuring elements $\{B_i^j\}$, shown in Fig. 5, are designed for this purpose [14]. Our procedure introduces more markers, where reliable edges are detected, and a few in homogeneous areas. This characteristic is desirable for SVO extraction: after the decision procedure, more accuracy in the segmentation *between* objects will be achieved, while keeping a small number of regions *inside* objects of approximately constant color.

Decision: We propose a joint distance measure criterion for the decision step that integrates color and edge information in the multivalued model

$$d(x, R_i^t) = \frac{1}{4} \{ |\overline{R}_i^t(\mathbf{R}) - \mathbf{R}^t(x)| + |\overline{R}_i^t(\mathbf{G}) - \mathbf{G}^t(x)| + |\overline{R}_i^t(\mathbf{B}) - \mathbf{B}^t(x)| + \mathbf{E}^t(x) \} \quad (6)$$

where $\overline{R}_i^t(\mathbf{R})$ represents the mean red value in the set R_i^t and equivalently for the other cases.

In the proposed watershed procedure, the more similar a given pixel x and its closest marker R_i^t are, the faster x is processed and assigned to R_i^t , unless x has been classified as an edge (reliable or not). Our distance criterion penalizes this case and delays the assignment of these pixels until the last stage of the decision process. Furthermore, when these pixels are delayed for classification, its neighbors are also delayed due to the watershed mechanism. This reduces the possibility that a given marker grows in a nondesired direction. In fact, our proposed distance criterion serves as a complement for the cases in which reliable edges are not extracted, so that markers are not defined.

To illustrate the single effect of the inclusion of intensity discontinuity information in the distance measure, we show in Fig. 6 the results obtained with a two-region synthetic image. Both regions have very similar color. The edge information is

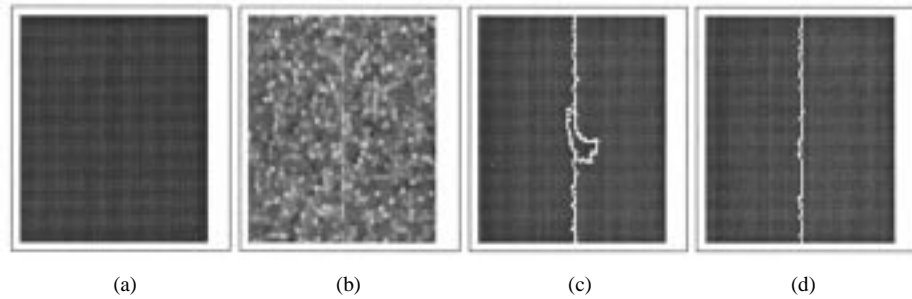


Fig. 6. (a) Two-region synthetic image. (b) Multiscale morphological gradient. (c) Segmentation using flat-zone marker extraction, color-only watershed and postprocessing. (d) Segmentation using flat-zone marker extraction, color and edges watershed, and postprocessing.

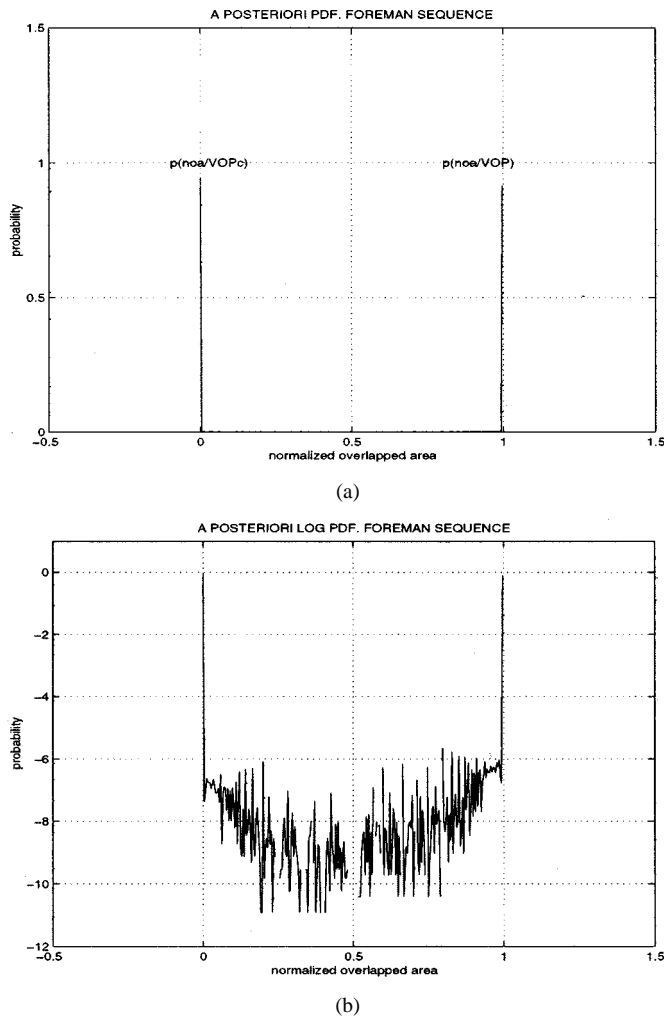


Fig. 7. (a) Modeling of the a posteriori distributions $p(noa|svo_0^t)$ and $p(noa|svo_j)$ for Foreman sequence. (b) Logarithm of these distributions.

obtained by a normalized multiscale morphological gradient operator [Fig. 6(b)] defined in [34]. We observe that the gradient image is noisy, although high values for the region boundary are detected. Markers are obtained by flat-zone detection, after simplification with a 3×3 structuring element. After the watershed, the partition postprocessing procedure (region merging) described in the next paragraph is applied. The segmentation result obtained with the use of color-only information is shown in Fig. 6(c), while the segmentation results generated with the joint color-edge distance measure is presented in Fig. 6(d). The

TABLE II
ML-ESTIMATED PARAMETERS FOR MPEG-4 SEQUENCES

Sequence	$\hat{\lambda}_1$	$\hat{\lambda}_2$	\hat{k}^0	\hat{T}_{noa}
Bream	142.21	113.89	2.44	0.45
Foreman	75.17	79.51	1.94	0.51
Hand	97.64	78.83	4.25	0.44

edge information prevented the marker located approximately at the center of the image from growing into the other side.

Postprocessing: After the watershed step, two morphological operations to reduce the final number of regions without affecting the accuracy in the segmentation are applied. In the first place, to remove small regions, an area connected operator is applied, in a way similar to [10]. We start by creating an image from P^t and Y^t by using \bar{Y}_i^t (mean intensity value) as region model. Then, the connected operator merges all regions smaller than a threshold with their closest neighboring regions of the largest size. The merging order is given by the size of the region. Finally, the flat-zone partition of the output image corresponds to the output partition. In the second place, to merge neighboring regions with the same color, another connected operator is applied. First, we create a color image from P^t and I^t , using $(\bar{R}_i^t, \bar{G}_i^t, \bar{B}_i^t)$ as the region model. The operator merges all neighboring regions that have the same color. Again, the flat-zone partition of the output image corresponds to the output partition.

As we will show in Section V, our proposed four-band segmentation method produces partitions with good characteristics for SVO extraction, and improves the results compared to the traditional watershed techniques. The obtained partitions will conform the basic domain on which the operators defined in the next section will be applied to generate video objects.

IV. PARTITION LATTICE OPERATORS FOR SEMANTIC VIDEO OBJECT EXTRACTION

As we stated in Section II, once a partition P^t is generated at each time t , the problem of SVO extraction becomes the *construction* of P_{SVO}^t from P^t and the temporal reference information TR . Remember that we have assigned the accurate extraction of region boundaries to the partition generation phase. As a result, the extraction operators $\{\psi^t\}$ can be thought of as a classification mechanism, that assigns each region $R_i^t \in P^t$ to the

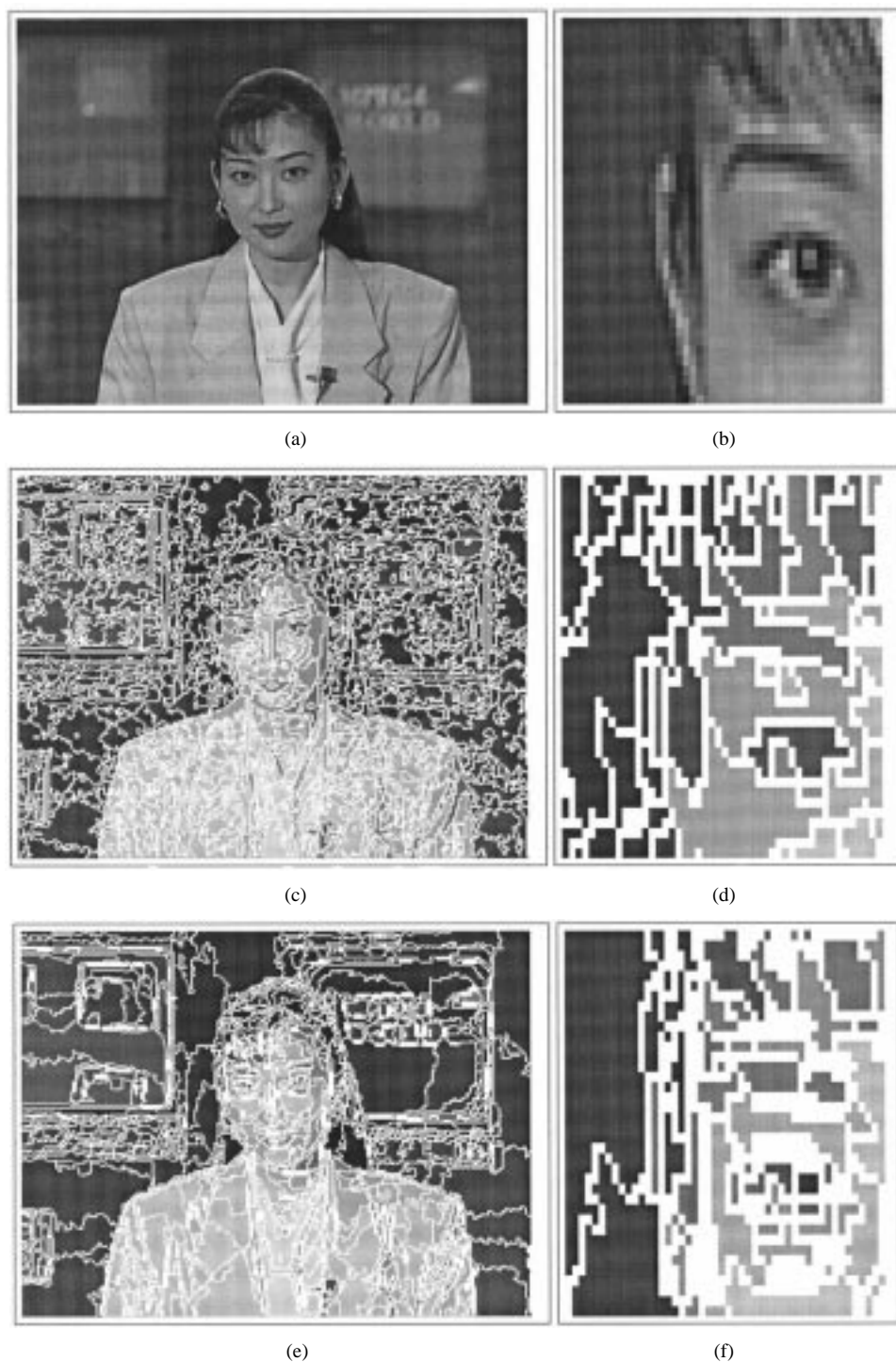


Fig. 8. Partition generation results. (a) *Akiyo* sequence. (b) Close-up of the scene for the ear/hair area. (c) Color-only traditional watershed segmentation. (d) Close-up of the same region. Background and foreground have been erroneously merged: artifacts in the SVO will be introduced. (e) Segmentation with our proposed approach. (f) Corresponding close-up. No erroneous merging has occurred.

appropriate SVO, so that no new spatial contours are introduced in the partition P_{SVO}^t . Statistical criteria can be systematically used in the definition of such operators.

Among a number of interesting theoretical properties of operators in complete lattices, three are particularly important in mathematical morphology: extensiveness, idempotence, and increasingness [14]. In our case, as the result of the region classification mechanism, the partitions P_{SVO}^t are produced in such a way that no new spatial contours are introduced in the extracted

SVOs. In fact, all possible operators that can be designed with this idea in mind satisfy one of the previously mentioned properties (the proofs for all the propositions in this paper can be found in the Appendix).

Property 1: The class of operators $\{\psi^t\}$ is extensive.

The proposed approach allows for the development of several extraction operators. In particular, both single and multiple-view operators are feasible to design under the same framework, and a variety of statistical criteria are available. Other morphological

concepts like connectivity can be also used in this framework. In this paper, we present one operator for the case in which the reference set is given by $TR = \{P_{SVO}^{t-1}\}$, and ψ is nontime-adaptive. The theory and results obtained for different cases will be reported in future publications.

A. Partition Operator Based On Regional Maximum Likelihood

The design of the partition operators can be formulated in terms of an optimality criterion to be satisfied. Note that, from the statistical point of view, SVO extraction (as has been formulated here) represents a process of assigning the regions of P^t to a given class, namely the objects in the scene; from the algebraic point of view, it corresponds to the design of extensive partition operators. We initially propose a partition operator $\psi_j(P^t)$ to construct each SVO_j^t from the partition P^t using the previous SVO partition (P_{SVO}^{t-1}) as reference

$$P_{SVO_j}^t = \psi_j(P^t)$$

where $P_{SVO_j}^t$ is the partition that divides the image support into the j th SVO and the rest of the scene. The generation of P_{SVO}^t is then straightforward.

Let $V_{\mathbf{I}^t, \mathbf{I}^{t-1}} : \mathcal{P}(\mathcal{E}) \rightarrow \mathcal{Z}^2$ be the mapping that computes a region motion vector, assuming a pure translational model, using the image sequence \mathbf{I}^t at times t and $t-1$, so that $V_i^t = V_{\mathbf{I}^t, \mathbf{I}^{t-1}}(R_i^t)$ denotes the motion vector computed for the region $R_i^t \in P^t$. Additionally, let $[X]_h$ represent the translated version of $X \subset \mathcal{Z}^2$ by $h \in \mathcal{Z}^2$: $[X]_h = \{x+h \mid x \in X\}$. The region attribute that will be used to construct the SVO partition at each time t , using the temporal reference $P_{SVO}^{t-1} = \{SVO_j^{t-1}, j \in \{1, \dots, M\}\}$ is defined as follows.

Definition 3: Given a partition P^t , and the SVO partition P_{SVO}^{t-1} , the normalized overlapped area between the i th region $R_i^t \in P^t$ and the j th SVO is given by

$$noa_{ij}^t = \frac{\text{card}([R_i^t]_{V_i^t} \cap SVO_j^{t-1})}{\text{card}(R_i^t)}. \quad (7)$$

This measure takes values between zero (no overlapping) and one ($R_i^t \subseteq SVO_j^{t-1}$), and will decide for the assignment of each region in P^t to the corresponding SVO. Obviously, each $R_i^t \in P^t$ belongs either to the j th SVO or to any other SVO in the scene depicted in the image sequence. In hypothesis-testing terms

$$H_0 : R_i^t \subseteq SVO_j^t \quad ; \quad H_1 : H_0^c. \quad (8)$$

The normalized overlapped area can be modeled as a continuous random variable \mathbf{noa} , taking values noa in $[0,1]$ (we drop the index t in what follows to simplify the notation). Let $svo_j, j = 1, \dots, M$ represent the j th possible class (i.e., the j th SVO), with prior probabilities $\Pr(svo_j)$, and let svo_j^c denote the set of all classes except the j th one, which implies $\Pr(svo_j^c) = 1 - \Pr(svo_j)$. With this setting, $\Pr(svo_j|noa)$ and $\Pr(svo_j^c|noa)$ represent the a posteriori conditional probabilities that corre-

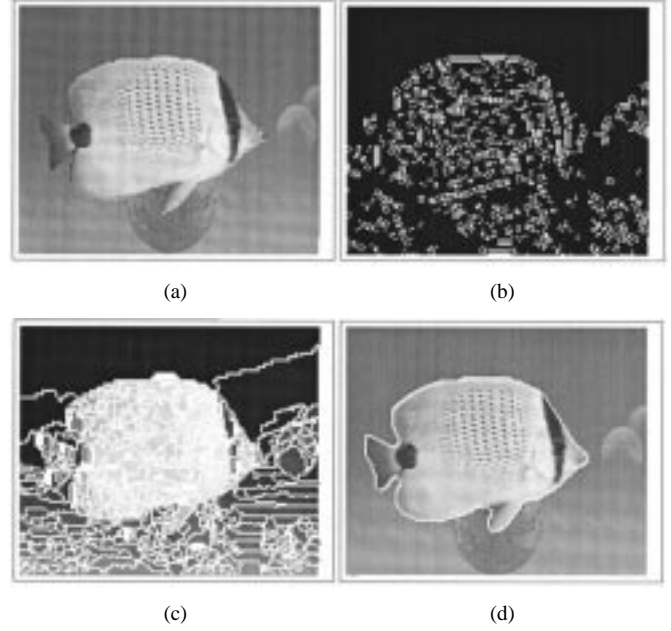


Fig. 9. (a) Bream sequence. (b) Edge-based markers. (c) Proposed segmentation. (d) Two SVOs: bream and background.

spond to H_0 and H_1 , respectively. We use the maximum a posteriori (MAP) criterion to map each region to an SVO [21]

$$\Pr(svo_j|noa) \underset{H_0}{\overset{H_1}{\leq}} \Pr(svo_j^c|noa) \quad (9)$$

such that the hypothesis H_x that is chosen is the one that has a larger a posteriori probability. Applying Bayes theorem on both sides of the expression and rearranging terms

$$\frac{p(noa|svo_j)}{p(noa|svo_j^c)} \underset{H_0}{\overset{H_1}{\leq}} \frac{\Pr(svo_j^c)}{\Pr(svo_j)} \quad (10)$$

where $p(noa|svo_j)$ represents the class-conditional probability density function. For the two-object case, we can assume equal priors ($\Pr(svo_j) = \Pr(svo_j^c)$), as foreground and background video objects may have any size and shape, and the expression reduces to the maximum likelihood criterion

$$L(noa) \equiv \frac{p(noa|svo_j)}{p(noa|svo_j^c)} \underset{H_0}{\overset{H_1}{\leq}} 1. \quad (11)$$

For the cases of larger number of objects, however, the exact expression is (10). Let k^t denote the ratio $\Pr(svo_j^c)/\Pr(svo_j)$. We propose to model the class-conditional probability density functions by exponential distributions

$$\begin{aligned} p(noa|svo_j^c) &= \lambda_1 e^{-\lambda_1 noa} u(noa) \\ p(noa|svo_j) &= \lambda_2 e^{-\lambda_2(1-noa)} u(1-noa) \end{aligned}$$

where $u(x)$ designates the step function. We believe that these distributions approximately model the real data: due to segmentation errors, $p(noa|svo_j)$ should be highly concentrated around $noa = 1$, and rapidly decay as $noa \rightarrow 0$. The dual situation holds for $p(noa|svo_j^c)$. In addition, the parameter values λ_i should make the conditional probabilities outside the interval



Fig. 10. SVO extraction results for two SVOs: *Akiyo* and *background*; *Akiyo* sequence. (a) SVO partition based on traditional watershed segmentation, flat-zone approach. Frames 20, 50, 60 and 70. (b) Partition based on traditional watershed segmentation, modified flat-zone approach, same frames. (c) Partition based on proposed approach.

$[0, 1]$ negligible. The problem has been reduced to finding an optimal threshold for noa

$$noa \underset{H_0}{\overset{H_1}{\leq}} \frac{\lambda_2 - \ln(\lambda_2/k^t \lambda_1)}{\lambda_1 + \lambda_2} = T_{noa}. \quad (12)$$

We can now write an expression for the proposed partition operator

$$P_{SVO_j}^t = \psi_j(P^t) = \{SVO_j^t, \mathcal{E} \setminus SVO_j^t\} \quad (13)$$

where $A \setminus B$ denotes set difference and

$$SVO_j^t = \bigcup_i R_i^t \text{ such that } noa_{ij}^t \geq T_{noa}. \quad (14)$$

The parameters λ_i and k^t can be estimated from the actual data. However, if we assume symmetry between the exponential distributions ($\lambda_1 = \lambda_2$), and $\lambda_i \gg k^t$, the expression for the optimal threshold can be further simplified and approximated as

$$T_{noa} = \frac{\lambda_2 - \ln(\lambda_2/\lambda_1)}{\lambda_1 + \lambda_2} + \frac{\ln k^t}{\lambda_1 + \lambda_2} \approx \frac{1}{2}. \quad (15)$$

This analysis shows that ψ_j , under the described assumptions, is equivalent to a tracking algorithm recently reported in [13] for the two-SVO case. It is interesting to notice that such an algorithm was based on a nonparametric classification technique, while ψ_j has been obtained by the MAP criterion, and some simplifying assumptions.

To extract the M SVOs present in the scene, ψ_j should be applied $M-1$ times (the M th SVO is always selected as the scene background). Finally, P_{SVO}^t can be directly generated from the set of partitions $\{\psi_j(P^t)\}_{j=1}^{M-1}$, by defining a partition operator $\psi_{RML}()$ for *regional maximum likelihood*

$$P_{SVO}^t = \psi_{RML}(P^t) = \bigwedge_{j=1}^{M-1} \psi_j(P^t) = \bigwedge_{j=1}^{M-1} P_{SVO_j}^t. \quad (16)$$

Two more important properties of ψ_{RML} can now be established [the same applies to ψ_j , as it is equivalent to ψ_{RML} for $M=2$ in (16)].

Property 2: ψ_{RML} is not increasing.

Property 3: ψ_{RML} is not idempotent. However, for situations of interest, it holds that $\psi_{RML}^2(P^t) = \psi_{RML}(P^t)$.

B. Statistical Validation

To justify our assumptions, we have performed statistical tests on several MPEG-4 video sequences. Fig. 7(a) shows the estimated conditional pdfs $p(noa|svo_j^c)$ and $p(noa|svo_j)$ for 200 frames of the *Foreman* sequence, where the SVO is the talking man. We can see that the first distribution is highly concentrated around zero, while the second one behaves similarly around one. In addition, both distributions are approximately symmetric. In Fig. 7(b), we display a log version of the distributions, to enhance the details. In fact, even though from the distributions we could be tempted to use thresholds in a large range, this is not convenient: every single misclassification will introduce artifacts in the extracted SVOs. We want to keep this situation to the minimum.

In Table II, we show the ML estimates for the parameters λ_i , for the two-SVO case (foreground object and background). Additionally, the priors $\Pr(svo_j)$ at each time t are estimated from the relative sizes of the SVOs at the previous frame of the video sequence, so that

$$\hat{k}^t = \frac{\Pr(svo_j^c)}{\Pr(svo_j)} \approx \frac{\text{card}(\mathcal{E} \setminus SVO_j^{t-1})}{\text{card}(SVO_j^{t-1})}. \quad (17)$$

Table II also shows the initial values of \hat{k}^t . It is observed that the assumption that $\lambda_i \gg k^t$ also holds, even for small objects, and that the estimated optimum threshold \hat{T}_{noa} is actually close to the approximated value. This fact validates the direct use of $1/2$ as the value of T_{noa} , which reduces the computational complexity. It is also pointed out that ψ_{RML} can tolerate SVO size changes.

V. RESULTS

In this section, we first show results that illustrate the performance of our multivalued watershed. Then, we present examples of the proposed framework for SVO extraction when it is included in the system mentioned in Section II.

A. Partition Generation Results

An example that illustrates the partition generation algorithm is shown in Fig. 8. One frame of the *Akiyo* sequence and a detail of a region difficult to segment are shown in Fig. 8(a) and (b). On one hand, the segmentation obtained with the traditional flat-zone approach, color-only segmentation is displayed in Fig. 8(c). The segmentation has done a good job in contrasted scene regions, but it has failed in some regions of similar color: the close-up in Fig. 8(d) indicates that background and foreground have been merged, due to weak intensity edges and absence of markers. On the other hand, after extracting edges, and edge-based markers, the partition that results from our proposed method is shown in Fig. 8(e). This segmentation has better characteristics for SVO extraction, as more real contours have been preserved. The corresponding close-up shows that the same regions have now been correctly segmented, as our method introduced markers on both sides of the detected weak edges.

As a second example, in Fig. 9 we show the results obtained with our methods for one image of the *Bream* sequence. Again, the generated partition has good accuracy for SVO extraction, as can be seen in Fig. 9(d).



Fig. 11. SVO extraction, *Foreman* sequence. The scene background has been removed for blue-screen emulation.

B. SVO Extraction

In this subsection we present the results obtained with our algorithms for different scenarios. To generate the first SVO partition, the user is allowed to specify the contours of each object using a GUI, following the technique described in [12]. The application of our methodology in the first frame deals with typical contour errors introduced by the user. Additionally, it is reasonable to assume that after any scene change, the original SVOs are not present in the scene. Therefore, the user is asked to input a new set of SVOs in case he/she is still interested in the extraction process.

We perform an objective evaluation of the spatial accuracy of those sequences for which a groundtruth is available. A criterion for evaluating the spatial distortion of an estimated

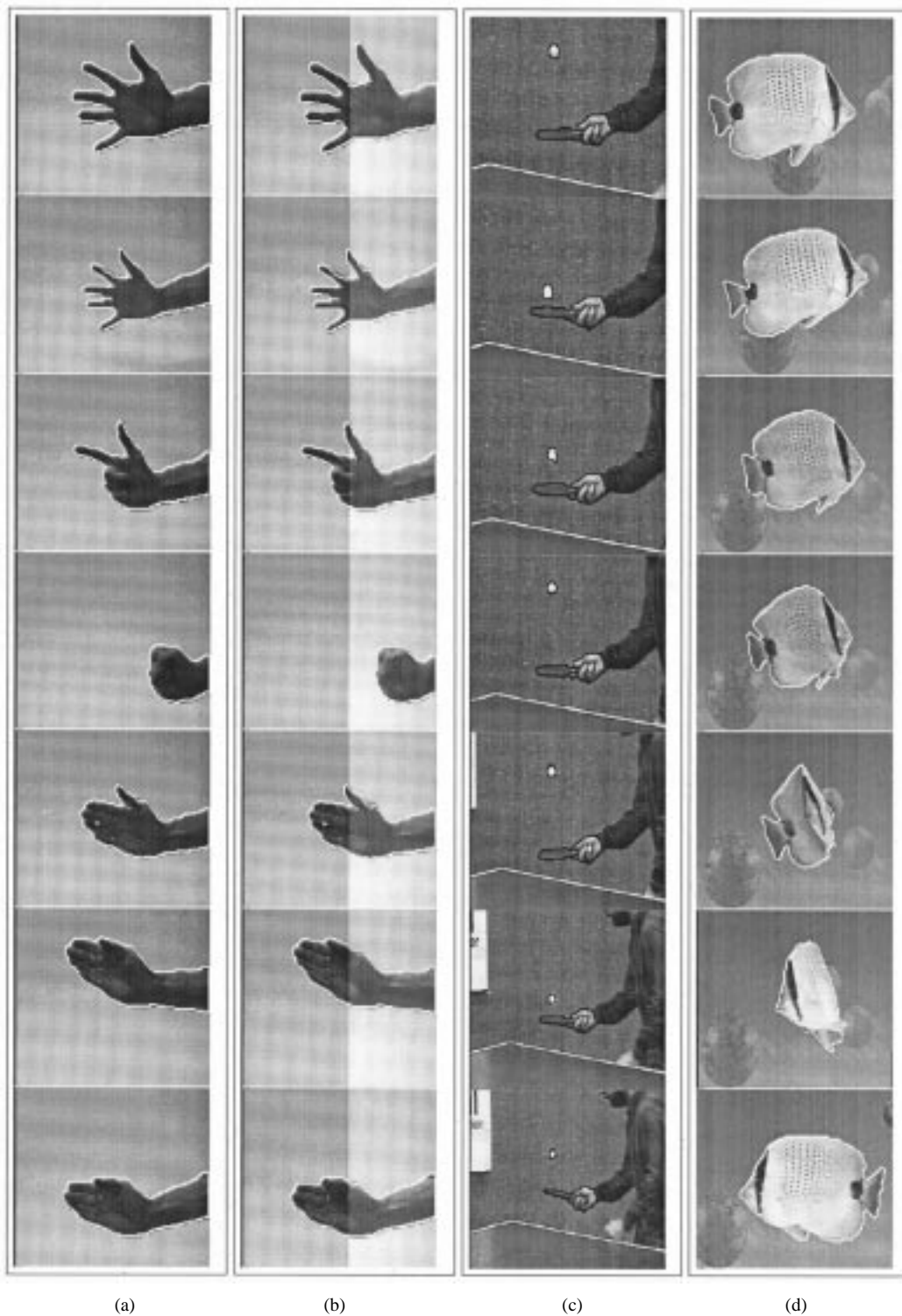


Fig. 12. SVO tracking. (a) *Hand* sequence. (b) Simulating illumination changes. (c) *Tennis* sequence. (d) *Bream* sequence. The scene background has not been removed to appreciate the scene characteristics.

VOP mask with respect to a reference mask at frame t is defined in [36]

$$d(P_{SVO_j}^t, P_{SVO_j}^t) = \frac{\sum_{x \in \mathcal{E}} (P_{SVO_j}^t(x) \text{ XOR } P_{SVO_j}^t(x))}{\sum_{x \in \mathcal{E}} P_{SVO_j}^t(x)} \quad (18)$$

where

$P_{SVO_j}^t$ j th estimated VOP mask, considered as a binary image;

$P_{SVO_j}^t$ corresponding reference VOP;

XOR denotes the exclusive-or logic operation.

1) *Static Camera, Small Object Motion*: SVO partitions for the *Akiyo* sequence obtained with the proposed partition lattice

operators and both the traditional watershed technique and the proposed segmentation are shown in Fig. 10. The sequence on the left side illustrates the case for a 3×3 flat-zone approach. We see that there are problems in areas of similar color; the tracking process will propagate these errors through time.

In the traditional approach, we could raise a question about the minimum size of the flat zones that should be considered as valid markers. The answer, however, is not clear, and the usual rule (the size of the structuring element used for simplification) is somewhat arbitrary. We have experimented with other rules, and Fig. 10(b) illustrates the quality of the extracted SVOs when a minimum fixed size of five pixels (central pixel plus its four closest neighbors) is used as threshold for marker selection. The figure shows again how the marker extraction phase impacts the final result.

Finally, in Fig. 10(c) we show the results obtained with our algorithm. We observe that the obtained VOP masks are more accurate and stable along time. Our visual judgement is supported by the objective evaluation provided in Fig. 13(a). As a reference of the obtained accuracy, we also show the distortion computed between the groundtruth and a 3×3 -eroded version of itself, which peels off the hand-made partition by approximately one pixel. It is important to highlight that for SVO extraction no objective measurements perfectly match human perception. For instance, compare frame 20 in Fig. 10(a) and (c), that are numerically similar in quality, but subjectively different, because even thin artifacts that merge background and foreground are visually disturbing. We are currently analyzing other objective quantities that reflect this fact.

2) *Moving Camera, Fast Motion*: The result with our proposed method for the *Foreman* sequence and two SVOs (*background* and *man*) is shown in Fig. 11. In this case, we have extracted the man from the scene and replace the background with a “blue screen.” We can see that the extracted SVO has high quality.

3) *Static Camera, Articulated Motion*: Fig. 12(a) shows the results obtained with a *Hand* sequence used for gesture recognition purposes [2]. This gray-level sequence presents fast, articulated motion and shades. We also show the results when we simulate a significant change of scene illumination, by introducing a large overexposed image region. Note that in both cases, the hand has been accurately segmented and tracked.

4) *Moving Camera, Multiple Rigid SVOs*: We illustrate this case in Fig. 12(c) with the *Tennis* sequence, in which the *ball* and *hand* SVOs have independent motion, while the scene experiences zooming. Both objects along with the background of the scene are correctly extracted.

5) *Moving Camera, Deformable Motion*: The result obtained with the *Bream* sequence is shown in Fig. 12(d). For this synthetic videoclip, we can also provide an objective evaluation. Fig. 13(b) shows the spatial distortion obtained for both the traditional flat-zone watershed and the variation in marker extraction discussed in this subsection, and for our proposed method; the latter extracts again more accurate SVOs.

The degree of accuracy of the SVO extraction process is bounded by two factors: 1) the resolution of the spatial segmentation and 2) the causal nature of the tracking process. On one hand, color similarity between different SVOs might

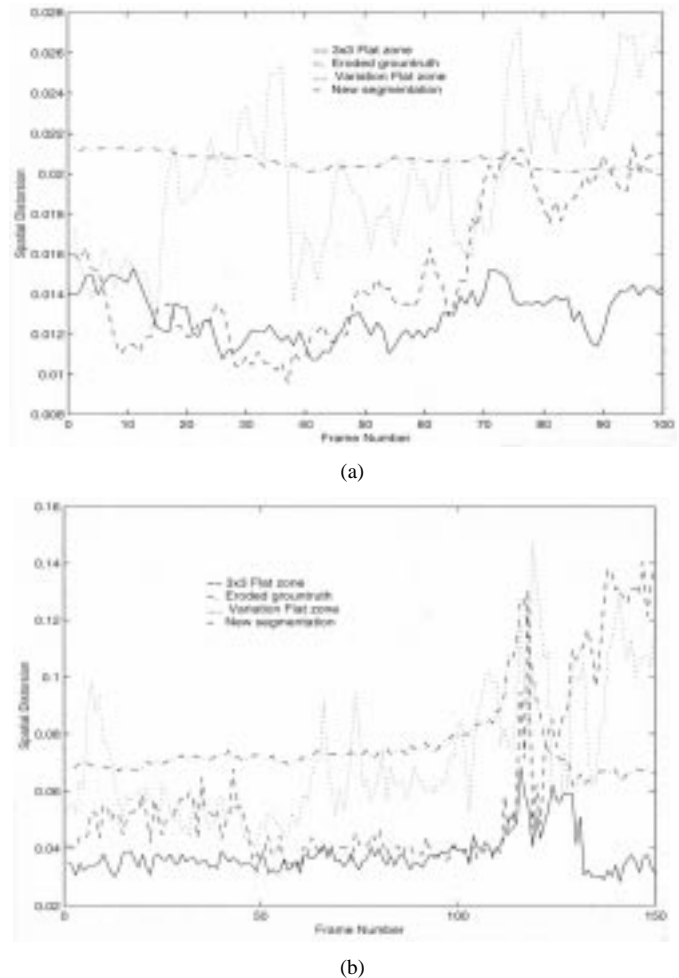


Fig. 13. Spatial distortion for traditional watershed methods and the proposed method for (a) *Akiyo* and (b) *Bream*.

result in poor edge detection, which can introduce segmentation errors in small areas, and originates the loss of small details (i.e., *Foreman*'s helmet). The resolution of the segmentation also degrades in highly cluttered scenes. On the other hand, the tracking process may propagate this loss of details in successive frames. Although the user-interactive approach provides a way of correcting such situations, we are currently investigating solutions for these open problems.

VI. CONCLUSION

We have described an approach for multiple SVO extraction based on spatial partition generation and extensive spatio-temporal operators on these partitions. The first stage has been addressed by developing a new multivalued morphological spatial segmentation that incorporates an edge-driven marker extraction method, and integrates color and edge information in the watershed algorithm. Our segmentation method naturally appeared as the consequence of searching for a method that could improve the contour localization of the traditional morphological segmentation techniques. Our statistical analysis of the watershed algorithm provided clues for such improvement. For the second step, the use of the partition lattice framework for SVO extraction is formal and general, as it allows for the modeling of

different schemes and may lead to the development of optimal algorithms. We have illustrated this approach by proposing a regional maximum likelihood operator. The integration of our algorithms into a SVO extraction system has produced SVO masks of improved quality compared to other traditional techniques. Experimental results for both synthetic and real video sequences have verified its effectiveness.

The formalization of the SVO extraction problem should allow for theoretical advances in this area as well as for the development of efficient interactive video authoring systems. These two directions have been included in our current lines of work.

APPENDIX

Definition 4: Given a space \mathcal{E} and its power set $\mathcal{P}(\mathcal{E})$, a *partition* of \mathcal{E} is a mapping $P : \mathcal{E} \rightarrow \mathcal{P}(\mathcal{E})$ such that $\forall x, y \in \mathcal{E}$, (i) $x \in P(x)$, and (2) $P(x) = P(y)$ or $P(x) \cap P(y) = \emptyset$. $P(x)$ is called the *zone* or *region* of P that contains x .

Definition 5: A *partition lattice* Π is a set of partitions $\{P_i\}$ of \mathcal{E} with a partial ordering relation (\leq), for which each of its subsets has an infimum and a supremum. It can be proved that the set of all partitions of \mathcal{E} constitutes a complete lattice.

Definition 6: The *partial ordering relationship* in a lattice of partitions is defined as $P_i \leq P_j \leftrightarrow P_i(x) \subseteq P_j(x), \forall x \in \mathcal{E}, P_i, P_j \in \Pi$. In this case, P_i is said to be *finer* than P_j .

Definition 7: The *infimum* of a set of partitions $\{P_i, i \in \mathcal{I}\}$ is defined as $(\bigwedge_i P_i)(x) = \bigcap_i P_i(x) \forall x \in \mathcal{E}$, i.e., it corresponds to the partition made of the intersections of all the regions in the original set of partitions. Additionally, the *supremum* of a set $\{P_i\}$ is given by $(\bigvee_i P_i)(x) = \bigcap \{B : B = \cup_i \cup_{y \in B} P_i(y), x \in B, B \in \mathcal{P}(\mathcal{E})\}$ which is the finest partition that is larger than each of the individual P_i . For the two-partition case, $(P_i \vee P_j)(x) = (P_i \vee P_j)(y)$ if $P_i(x) = P_i(y)$ or $P_j(x) = P_j(y)$.

Definition 8: The *least* and *greatest* elements of Π correspond to the finest partition P_O and the coarsest partition P_I , such that $P_O(x) = x$ and $P_I(x) = \mathcal{E}$ for all $x \in \mathcal{E}$.

For purposes of indexing of the zones of a partition, it is convenient to use the following notation: $P = \{R_i, i \in \mathcal{I}\}$ where $R_i = \cup x \in \mathcal{E}$ such that $P(x) = R_i$

Property 1: The proposed class of operators $\{\psi^t\}$ is extensive.

Proof: A lattice operator ψ is extensive iff $P^t \leq \psi(P^t) \forall P^t \in \Pi$. The proof directly follows from (1) and (2) \square

Property 2: ψ_{RML} is not increasing.

Proof: An operator ψ is increasing iff $\forall P_i^t, P_j^t$ such that $P_i^t \leq P_j^t, \psi(P_i^t) \leq \psi(P_j^t)$ (the ordering relation is preserved). It is simple to show that this is not the case with a counterexample. Let P_i^t be a partition of \mathcal{E} that consists of three regions, labeled B (background), H (head), and S (shoulders), respectively. Let P_j^t be another partition that consists of two regions, H and $E = B \cup S$ (erroneously merged regions). By construction, $P_i^t \leq P_j^t$. Additionally, let us assume that there is no motion and that P_{SVO}^{t-1} is correctly composed of the background B and the object $O = H \cup S$. Applying ψ_{RML} to the two partitions, we obtain $\psi_{\text{RML}}(P_i^t) = P_{\text{SVO}}^{t-1}$, and $\psi_{\text{RML}}(P_j^t) = P_j^t$, but obviously P_{SVO}^{t-1} is not finer than P_j^t , so ψ_{RML} is not increasing. \square

Similarly, it can be proved that ψ_{RML} is not decreasing.

Property 3: ψ_{RML} is not idempotent. However, for situations of interest, it holds that $\psi_{\text{RML}}^2(P^t) = \psi_{\text{RML}}(P^t)$.

An operator ψ is idempotent iff $\psi(\psi(P^t)) = \psi^2(P^t) = \psi(P^t)$; for all partitions P^t in the lattice Π . Even though this situation might not be true for *all* partitions, it holds for many practical situations. In particular, as we stated earlier in the paper, the result of applying ψ_{RML} depends on temporal information by the following mechanism:

- 1) region motion estimation and compensation between two consecutive frames;
- 2) computation of *noa* based on the estimated motion vectors and the previous SVO partition P_{SVO}^{t-1} . By definition, P_{SVO}^{t-1} is a partition with a very small number of regions. Each region in P_{SVO}^{t-1} represents one SVO (unless the SVO is disconnected). The largest region normally corresponds to the *scene background SVO*.

The regions of P^t are then classified, and form a new partition. Let denote this partition by $P_X^t = \psi_{\text{RML}}(P^t)$. P_X^t is also composed of a very small number of regions, usually equal to the number of regions of P_{SVO}^{t-1} (unless some object disconnection has occurred in this frame).

Now, a second application of the operator classifies each of the regions in P_X^t repeating the procedure just described. The motion of the regions in the original partition P^t is, in general, not equal to the motion of the merged regions in P_X^t (that describes the match of each SVO). However, the regions in P_X^t will be still mostly overlapped with the appropriate SVO in P_{SVO}^{t-1} . As a result, the reclassification simply assigns each region in P_X^t to the SVO whom they were already assigned after the first application of the operator.

We have tested this property in all the video sequences reported in this paper (i.e., we have applied the operator twice on each P^t). We have observed that the second application of the operator does not have any effect. In other words, we have verified that $\psi_{\text{RML}}(P_X^t) = \psi_{\text{RML}}(P^t) = P_X^t$. Therefore, we define the SVO partition at time t as $P_{\text{SVO}}^t = P_X^t$. \square

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their effort to improve the quality of this paper.

REFERENCES

- [1] S. Beucher and F. Meyer, "The Morphological Approach to Segmentation: The Watershed Transformation," in *Mathematical Morphology in Image Processing*, E. Dougherty, Ed. New York: Marcel Dekker, 1993, pp. 433–481.
- [2] M. J. Black and A. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Comput. Vis.*, vol. 26, no. 1, pp. 63–84, 1998.
- [3] A. Blake and M. Isard, *Active Contours*. New York: Springer-Verlag, 1988.
- [4] F. Bremond and M. Thonnat, "A context representation for surveillance systems," in *Proc. Workshop Conceptual Descriptions from Images at the Eur. Conf. Computer Vision (ECCV)*, Cambridge, U.K., Apr. 1996.
- [5] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 679–698, Nov 1986.
- [6] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "A Fully Automated Content-Based Video Search Engine Supporting Spatiotemporal Queries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 602–615, Sept. 1998.

- [7] J. G. Choi, M. Kim, M. H. Lee, C. Ahn, S. Colonnese, U. Mascia, G. Russo, P. Talone, R. Mech, and M. Wollborn, "Combined algorithm of ETRI, FUB and UH on core experiment N2 for automatic segmentation algorithm of moving objects," Stockholm, Sweden, ISO/IEC JTC1/SC29/WG11/MPEG97/m2383, 1997.
- [8] P. Correia and F. Pereira, "The role of analysis in content-based video coding and indexing," *Signal Processing*, vol. 66, pp. 125–142, Apr. 1998.
- [9] V. Garcia-Garduno, "Une approche de compression orientee-objets par suivi de segmentation basee mouvement," Ph.D. dissertation, Univ. de Rennes I, Rennes, France, 1995.
- [10] L. Garrido, P. Salembier, and D. Garcia, "Extensive operators in partition lattices for image sequence analysis," *Signal Processing*, vol. 66, no. 2, pp. 157–180, Apr. 1998.
- [11] C. Gu, "Multivalued Morphology and segmentation based coding," Ph.D. dissertation, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, 1995.
- [12] C. Gu and M.-C. Lee, "Semiautomatic Segmentation and Tracking of Semantic Video Objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 525–538, Sept. 1998.
- [13] —, "Semantic Video Object Tracking Using Region-based Classification," in *Proc. IEEE ICIP*, Chicago, IL, Oct. 1998, pp. 643–647.
- [14] H. J. A. M. Heijmans, *Morphological Image Operators*. New York: Academic, 1994.
- [15] "Special Issue on Segmentation, Description, and Retrieval of Video Content," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 521–696, Sept. 1998.
- [16] S. Z. Li, *Markov Random Field Modeling in Computer Vision*. New York: Springer-Verlag, 1995.
- [17] F. Marques and J. Llach, "Tracking of generic objects for video object generation," in *Proc. IEEE ICIP*, Chicago, IL, Oct. 1998, pp. 628–632.
- [18] T. Meier and K. N. Ngan, "Automatic Segmentation of Moving Objects for Video Object Plane Generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 525–538, Sept. 1998.
- [19] F. Meyer and P. Boutheymy, "Region-based tracking using affine motion models in long image sequences," *CVGIP: Image Understanding*, vol. 60, pp. 119–140, 1994.
- [20] F. Meyer, "Minimum spanning forests for morphological segmentation," in *Mathematical Morphology and its Applications to Image Processing*, J. Serra and P. Soille, Eds. Norwell, MA: Kluwer, 1994.
- [21] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1993.
- [22] J. B. T. M. Roerdink and A. Meijster, "The Watershed Transform: Definitions, Algorithms and Parallelization Techniques," Institute for Mathematics and Computer Science, University of Groningen, Groningen, The Netherlands, IWI 99–9-06, 1999.
- [23] W. Rucklidge, *Efficient Visual Recognition Using the Hausdorff Distance*. New York: Springer-Verlag, 1996.
- [24] E. Saber, A. M. Tekalp, and G. Bozdagi, "Fusion of color and edge information for improved segmentation and edge linking," *Image Vis. Comput.*, vol. 15, no. 10, pp. 769–780, Oct. 1997.
- [25] P. Salembier, "Morphological Multiscale segmentation for image coding," *Signal Processing*, vol. 38, no. 3, pp. 359–386, Sept. 1994.
- [26] P. Salembier and M. Pardas, "Hierarchical segmentation for image sequence coding," *IEEE Trans. Image Processing*, vol. 3, pp. 639–651, Sept. 1994.
- [27] J. Serra, "Image Analysis and Mathematical Morphology," in *Theoretical Advances*. New York: Academic, 1988, vol. II.
- [28] —, "Anamorphoses and function lattices (multivalued morphology)," in *Mathematical Morphology in Image Processing*, E. R. Dougherty, Ed. New York: Marcel Dekker, 1993, pp. 483–523.
- [29] J. C. Simon, *Patterns and Operators: the Foundations of Data Representation*. Oxford, U.K.: North Oxford Academic, 1986.
- [30] D. S. Sivia, *Data Analysis. A Bayesian Tutorial*. London, U.K.: Oxford Univ. Press, 1996.
- [31] C. Toklu, A. T. Erdem, and A. M. Tekalp, "2D mesh-based synthetic transfiguration of an object with occlusion," in *Proc. IEEE ICASSP*, Munich, Germany, 1997, pp. 2649–2652.
- [32] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE CVPR*, Maui, HI, June 1991, pp. 586–591.
- [33] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 2, pp. 176–201, Apr. 1993.
- [34] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 540–546, Sept. 1998.
- [35] H. Wang and M. Brady, "Real-time corner detection algorithm for motion estimation," *Image Vis. Comput.*, vol. 13, pp. 695–703, Nov. 1995.
- [36] M. Wollborn and R. Mech, "Refined procedure for objective evaluation of video object generation algorithms," Doc. ISO/IEC JTC1/SC29/WG11 M3448, 1998.

Daniel Gatica-Perez (S'01) received the B.S. degree in electronics engineering from the University of Puebla, Mexico, in 1993, and the M.S. degree in electrical engineering from the National University of Mexico in 1996. He is currently working toward the Ph.D. degree in electrical engineering at the University of Washington at Seattle.

His research interests include video analysis, mathematical morphology, and statistical pattern recognition.

Chuang Gu (M'97) received the B.S. and M.S. degrees in computer science from Fudan University, Shanghai, China, in 1986 and 1989, respectively, and the Ph.D. degree in electrical engineering from Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1995. His Ph.D. dissertation was on multivalued morphology, segmentation-based coding, and tracking.

From 1989 to 1990, he was an Assistant Director at the CAD center of Fudan University. In 1991, he was an Associate Research Fellow at the European Organization for High Energy Physics (CERN). During 1992–1995, he was with the Signal Processing Laboratory, EPFL, as a Research Assistant. Since 1995, he has been with Microsoft Corporation, Redmond, WA, carrying out research and development on video processing, analysis and coding for Windows Media Technology (available at <http://www.microsoft.com/windows/windowsmedia/>). He has published more than 40 papers and earned six patents on these subjects. Since 1999, he has been an Affiliate Professor at the University of Washington.

Dr. Gu has served as an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA since 1999. He received the Young Investigator Award from SPIE Visual Communication and Image Processing in 1999.

Ming-Ting Sun (S'79–M'81–SM'89–F'96) received the B.S. degree from National Taiwan University in 1976, the M.S. degree from University of Texas at Arlington in 1981, and the Ph.D. degree from University of California at Los Angeles in 1985, all in electrical engineering.

He joined the University of Washington at Seattle in August 1996, where he is a Professor. Previously, he was the Director of the Video Signal Processing Research Group, Bellcore, Red Bank, NJ. He has been awarded seven patents and has published more than 100 technical papers, including ten book chapters in the area of video technology.

Dr. Sun is currently the Editor-in-Chief of IEEE TRANSACTIONS ON MULTIMEDIA, and was prior Editor-in-Chief of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) from 1995 to 1997. He was a conference General Co-Chair of the Visual Communications and Image Processing 2000 Conference (VCIP2000). From 1988 to 1991, he served as the Chairman of the IEEE Circuits and Systems (CAS) Standards Committee and established an IEEE Inverse Discrete Cosine Transform Standard. He received a Golden Jubilee Medal from the IEEE CAS Society in 2000, was a co-recipient of the TCSVT Best Paper Award in 1993, and received an Award of Excellence from Bellcore in 1987 for the work on Digital Subscriber Line.