

Multiview Extensive Partition Operators for Semantic Video Object Extraction

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

Abstract—Occlusion/disocclusion is one of the fundamental problems for semantic video object (SVO) extraction, where pixel-wise accuracy is required. This issue is critical because the degradation in tracking due to object occlusion/disocclusion significantly increases the amount of user interaction required in off-line video editing applications. In this paper, we present an approach based on the application of an extensive operator on a lattice of partitions, which exploits information from various views of the scene, based on a probabilistic formulation. Our multiview operator builds on the regional application of the Maximum a Posteriori principle, by integrating a single-view region classification stage with a multiview stage that improves classification for those disoccluded regions labeled as uncertain. Results on several real sequences show that our approach improves the SVO tracking compared to the single-view case and that, as a result, increases the quality of the extracted SVOs and reduces the total amount of user interaction.

Index Terms—Mathematical morphology, multiview partition operators, object tracking, semantic video object extraction.

I. INTRODUCTION

THE UNDERSTANDING and characterization of the dynamics of the objects in a scene constitute two of the most useful tasks in video analysis. Real-world video objects often experience motion transformations, illumination changes, scene disappearance and reappearance, and deformations. As time evolves, they usually do not remain invariant with respect to any low-level vision feature. Furthermore, objects can change their appearance either partly or completely within a few frames (for example, when opening a book of dark cover, or taking one's sweater off). Some of the latter situations are the result of large object (self) occlusion and disocclusion and are fundamental problems in object tracking [3].

In particular, extraction of semantic video objects (SVOs) is a necessary step for the wide applicability of the next generation of multimedia standard (MPEG-4) [19], and it poses two demanding requirements: 1) SVOs should be allowed to be arbitrary (objects represent human abstractions) and 2) the accuracy in extraction should be pixel-wise. Because of the complexity

of the problem, several researchers have proposed the use of semiautomatic methods to fulfill such impositions [4]–[6], [11], [12], [15], [17], [20], [25]. However, even though some ways of human-machine interaction appear to be evident (specification of the SVOs at the first frame of a video clip, or restarting of the system when the extraction quality becomes poor) [11], the devising of more efficient mechanisms of interaction still constitutes an open question.

Inscribed in the complete lattice framework proposed in mathematical morphology, we have recently presented a methodology for SVO extraction based on a 2-D region representation that consists of the generation of accurate spatial partitions followed by the application of extensive extraction operators on such partitions. Specifically, we introduced a regional maximum likelihood operator to implement single-view SVO tracking [9]. The extraction results generated by this approach have good quality, as the partition operators can handle shape deformations, partial disocclusion, change of illumination, and the presence of multiple objects, in moderate clutter. However, we believe that in order to deal with large object disocclusions and sudden appearance changes (including object reappearance in the scene), richer information of what each SVO represents at different instances, i.e., multiple views of the scene, should be provided to improve the tracking process, and subsequently, to increase the quality of the extracted objects. As a result of this investigation, in this paper we present an approach based on the design of a multiview extensive operator on a lattice of partitions for SVO tracking purposes. It uses information from a set of user-generated SVO partitions of the scene, provided at the beginning of the extraction process. The formulation of operators on lattices of partitions naturally allows for the definition of multiple view schemes. Our operator is derived from a regional probabilistic formulation and the application of the Maximum a Posteriori (MAP) principle to two region classification cases: one that relies on short-term (single-view) tracking information, and another one that uses information from multiple instances to solve for disoccluded regions when the single-view mechanism detects uncertain regions. We show that the multiview operator improves the SVO tracking compared to the single-view case, which results in an increase of quality of the extracted SVOs (measured both objectively and subjectively) and a reduction of the average amount of user interaction needed to perform non real-time SVO extraction.

The rest of the paper is organized as follows. Section II describes the formulation of the problem and discusses some previous works in this field. Section III presents our multiview extensive partition operator. Section IV shows some results ob-

Manuscript received November 1999; revised February 2001. The work of D. Gatica-Perez was supported by the Fulbright-CONACyT Ph.D. scholarship program, and by the National University of Mexico. This paper was recommended by Associate Editor S.-F. Chang.

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

M.-T. Sun is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA.

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

Publisher Item Identifier S 1051-8215(01)05278-8.

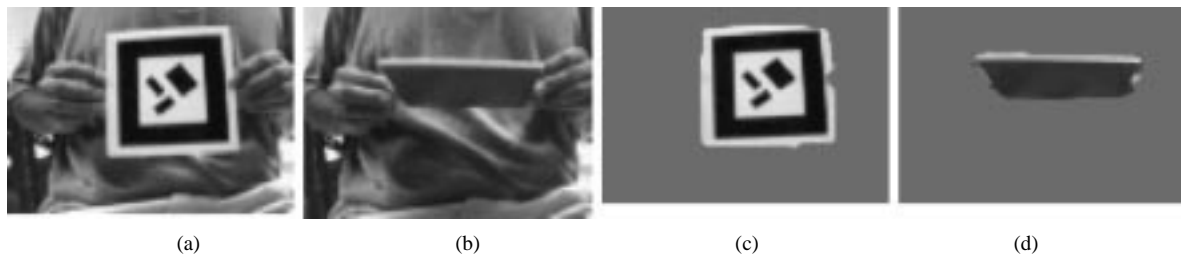


Fig. 1. *Square* test sequence. The *Square* SVO has two very different appearances: (a) front (frame 0) and (b) back (frame 68). The user defines the SVO partitions for these frames (c) and (d).

tained with typical test sequences when the operator is integrated in an SVO extraction system and discusses its performance. Section V provides some conclusions and future research directions.

II. FORMULATION OF THE PROBLEM

As described in [9], a complete lattice of partitions constitutes an algebraic structure in which the SVO extraction problem can be analyzed [22] (refer to the Appendix, where the notations and definitions used in this paper are provided). Our methodology constitutes a special case of the complete lattice approach in mathematical morphology [22], [13]. A few works have been previously developed using this framework, for simple segmentation problems [22], and to point out connections between connected operators and region merging algorithms [10]. We have focused its usage for SVO extraction.

A lattice both defines the notion of partial order relation between the elements of a set [7] and allows for the study of properties of the operators that can be defined on it [13]. A complete lattice of partitions Π (Appendix, Definition 2) of the domain E of a multivalued image sequence \mathbf{I}^t consists of all the possible partitions of E , spanning from the finest partition, in which each pixel of an image constitutes a region, to the coarsest one, that includes all the image pixels in one single region (Appendix, Definition 5). A lattice of partitions establishes the basis to construct a hierarchy of segmentations [10]. We have proposed that SVO extraction can be addressed with a two-step methodology, iterated at each time t in the video sequence.

- 1) Generation of accurate, fine spatial partitions $P^t = \{R_i^t, i \in \{1, \dots, N^t\}\}$, where R_i^t denotes the i th region, and N^t indicates their total number at time t that do not depend on noisy motion information, and hence preserve the real object contours.
- 2) Design of operators $\{\psi(\cdot)\}$ on a lattice of partitions Π (Appendix, Definition 8), based on statistical information such that, once applied on the partitions P^t , they generate a partition of SVOs, $P_{\text{SVO}}^t = \{\text{SVO}_j^t, j \in \{1, \dots, M\}\}$, where SVO_j^t denotes the j th SVO, and M is the number of SVOs in the scene (Appendix, Definition 7), without introducing any new spatial contours. Operators with this property are called *extensive* (Appendix, Definition 11).

In our scheme, *semantics* is introduced by the user, who specifies a *reference SVO partition set* (Appendix, Definition 8) denoted by TR , and composed of K SVO partitions of the scene

at different time instances, i.e., object partitions that correspond to various scene views

$$TR = \{P_{\text{SVO}}^{t_k}, k \in \{1, \dots, K\}\} \quad (1)$$

such that

$$P_{\text{SVO}}^t = \psi_{TR}(P^t). \quad (2)$$

This framework formalizes the concept of region-based SVO extraction, allows for the definition of operators, discloses some of their properties, and integrates the user information into the procedure [9]. In (1), when $K = 1$ and $t_K = t - 1$, the partition reference set is given by $TR = \{P_{\text{SVO}}^{t-1}\}$, which means that the generation of the current SVO partition will use only the previous object partition as a reference. The SVO extraction process reduces to the single-view extraction case described in [9]. This approach naturally supports the design of multiple view extractor operators, when $K > 1$. In particular, for video composition and editing applications in which the entire video clips are available in advance, user interaction in a multiview approach consists of two steps: 1) *off-line* selection of a number of *key frames* that reflect the different appearances of the SVOs and 2) generation of the corresponding SVO partitions to construct the set TR . Fig. 1(a) and (b) show a *Square* sequence where an object with two different facets is rotated in front of a camera. When a user—who naturally knows that his/her object of interest has more than one appearance—defines the reference SVO partition set as shown in Fig. 1(c) and (d), the introduction of semantics in the system makes it much richer than any single-view scheme. The initial SVO partitions are defined through the specification of the objects' outline, following an efficient procedure proposed in [11].

Multiview approaches have been used in the past for object recognition (eigenfaces [23] and 3-D objects in uncluttered background [18]) and tracking purposes (deformable shape models or eigenshapes [1], and eigentracking [2]). The common idea of all of these methods consists in generating a compact representation of the target data, by using a number of training examples. Several variations of principal component analysis (PCA) are performed to obtain such representation, which is then used to perform the matching/tracking tasks in the eigen-domain. There exist some important differences between these techniques. The method in [2] uses a rectangular region as the representation of the object support, while the one in [1] employs a parameterized object contour. Additionally, the method in [1] uses different objects with the same *shape variations* (walking people) to learn the shape models, while

gray-level appearance of objects of a same class (soda cans) are used for training in [2]. Finally, one work that deserves a special mention is the one reported in [3], which constitutes a long-term study on probabilistic modeling of objects via active contours. For all the described techniques, the results are encouraging for several real situations, e.g., shading, pose variations, and cluttered background. However, these methods do not explicitly deal with the pixel-wise precision that is required for extraction of arbitrary, deformable SVOs. Furthermore, for personal video editing applications, users can define anything that is present in the scene as an SVO of interest (a human body, an arm, a hand, or a finger). This represents limitations for the applicability of some of these techniques because training samples of arbitrary objects might not be easily available. Therefore, we believe that the described constraints point out to a matching strategy in the image domain. In this line of work, one recent study based on a deformable contour SVO representation is reported in [15], in which starting from two initial, user-provided object contours, any contour in intermediate frames can be interpolated. The interpolation is based on: 1) two-directional (forward, backward) contour tracking procedures that generate a pair of candidate contours and 2) a merging phase to produce one single contour per frame. Our methodology only coincides with the previous approaches in the use of multiple information for the SVO definition. As opposed to [15], we have adopted a 2-D region-based SVO structure representation, and formulated the problem of SVO extraction as an issue of designing morphological extensive operators on lattice of partitions. As we mentioned before, both single-view and multiple view operators can be inscribed in the same framework.

III. MULTIVIEW EXTENSIVE PARTITION OPERATORS

A. Formulation

In our methodology, once an accurate partition P^t is generated, the problem becomes the construction of the partitions of SVOs, P_{SVO}^t , based on P^t , the reference SVO partition set TR , and statistical criteria. As the definition of accurate object borders has been assigned to the partition generation step, the SVO extraction operators $\{\psi(\cdot)\}$ are thought of as a classification mechanism which assigns each region in the current spatial partition $R_i^t \in P^t$ to the appropriate SVO, without creating new contours. It is well known that the performance of any classification algorithm highly relies on a wise feature selection [8].

The design of the operators can be formulated in terms of optimality criteria. In this section, we first present the extraction operator for the two-view case ($K = 2$). The generalization to more views is straightforward.

Let ψ_j denote the operator that extracts the j th SVO, i.e., that partitions the image domain E into the j th SVO and the rest of the scene, such that $P_{\text{SVO}_j}^t = \psi_j(P^t)$. At each instant, each region $R_i^t \in P^t$ belongs either to such SVO or to any other SVO. In standard binary hypothesis testing notation

$$H_0: R_i^t \subseteq \text{SVO}_j^t; \quad H_1: H_0^c. \quad (3)$$

Let svo_j , $j = 1, \dots, M$ represent the j th possible class (i.e., the j th SVO), with prior probability $\Pr(\text{svo}_j)$, and let

svo_j^c denote the set of all classes except the j th one, which implies that $\Pr(\text{svo}_j^c) = 1 - \Pr(\text{svo}_j)$. Assume that the two keyframes, selected at the beginning of the extraction process, correspond to instances zero and t_v . Additionally, for all subsequent frames ($t \geq 1$), we define the reference SVO partition set to be composed of the SVO partitions at the previous frame, and at the keyframe at time t_v , i.e., $TR = \{P_{\text{SVO}}^{t-1}, P_{\text{SVO}}^{t_v}\}$. Furthermore, two spatio-temporal features are extracted for every region and for each of the views: the *normalized overlapped area* $\text{noa}_{ij}^{t_k}$ (Appendix, Definition 9), and the corresponding *normalized matching error* $e_i^{t_k}$ (Appendix, Definition 10). Both features are computed based on the motion vector estimated for each region, using the current frame and the corresponding keyframe, and assuming a translational model. More complex motion models could be used at an increased computational expense.

On one hand, the normalized overlapped area is a measure of the matching of each R_i^t with the j th SVO at time t_k . In real scenes, with objects experimenting nonrigid motion and illumination changes, a single motion vector that describes the correspondence of a region of arbitrary shape is likely to include errors. However, the computation of the normalized overlapped area filters out much of the uncertainty present in motion information because it indicates whether most of the region is overlapped with a certain object, rather than indicating where the exact match occurs. On the other hand, the normalized matching error is a measure of the matching of each region in the current partition with the scene at time t_k . Both quantities can take values between zero and one. Fig. 2 illustrates the feature extraction process.

Let noa^{t-1} and noa^{t_v} be the continuous random variables that represent the normalized overlapped area between R_i^t and the SVO partitions at times $t-1$ and t_v . Let e^{t-1} and e^{t_v} be the corresponding normalized matching errors. In a regional MAP formulation, we would like to maximize the conditional probability of assigning each region to the correct SVO [21], [8] as follows:

$$\text{svo}^* = \arg \max \Pr(\text{svo}_j | \text{noa}^{t-1}, \text{noa}^{t_v}, e^{t-1}, e^{t_v}) \quad (4)$$

or equivalently minimize the probability of misclassification. For the single-view case, in which we only rely on information from the previous frame, the above equation can be simplified to

$$\text{svo}^* = \arg \max \Pr(\text{svo}_j | \text{noa}^{t-1}) \quad (5)$$

which represents the original formulation proposed in [9]. Indeed, we showed that the operator obtained from the maximization of (5) is regionally optimal, assuming that for each region R_i^t there exists a corresponding region in the previous frame, so that a valid normalized overlapped area can be computed. In fact, this assumption is reasonable and works adequately for many scenarios (including deformable motion, multiple objects, and global motion in moderate clutter), with exception of those cases in which sudden disoccluded regions are quite different from the surrounding existing SVOs. However, the normalized matching error can be used as a measure of the likelihood of each region to be present in a given view. Most of the regions in

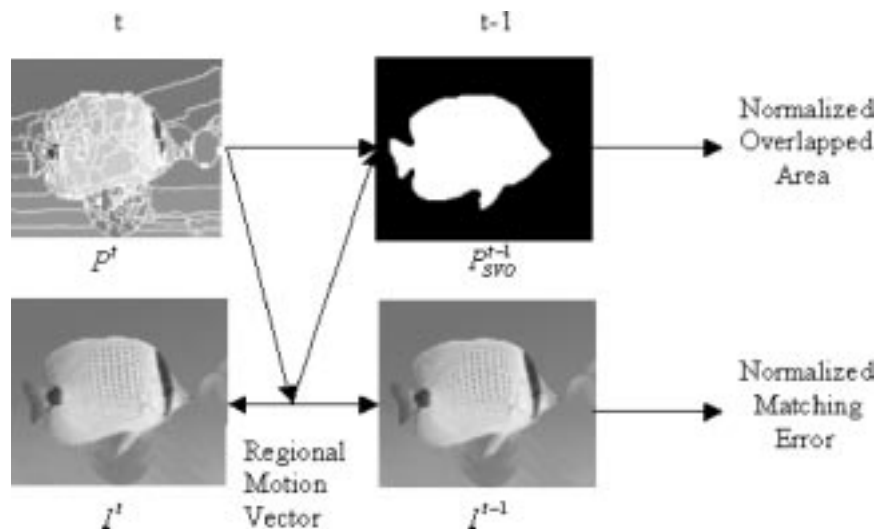


Fig. 2. Feature extraction process. The normalized overlapped area and the normalized matching error are computed for every region at each time, and used for classification.

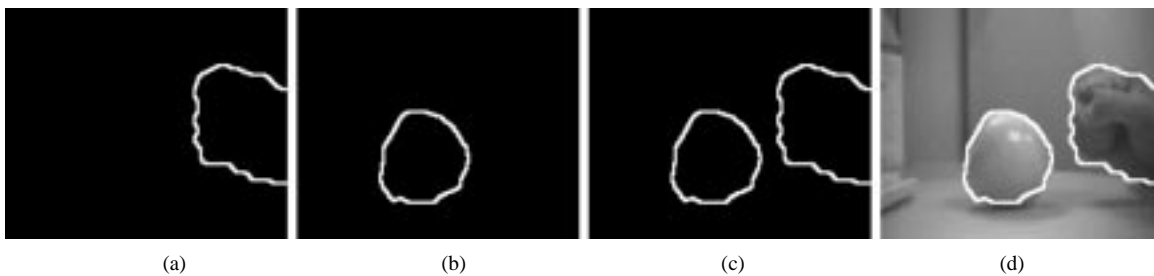


Fig. 3. The infimum of a set of partitions corresponds to the partition made of the intersections of all the regions in the original set. (a) P_1 ; (b) P_2 ; (c) infimum of P_1 and P_2 ; (d) extracted SVOs.

a partition have small matching error values (see Fig. 3), with exception of highly textured and disoccluded regions that are significantly different from the surrounding SVOs (called *uncertain regions* in the following). Situations like object reappearance in the scene, or sudden change of appearance fit in this category. In those cases, the single-view formulation may introduce region misclassifications. Nevertheless, these are exactly the cases for which multiple views of the scene can offer richer information for classification.

The multiview approach could be formulated as a global matching problem between all the regions in each partition and all the scene views. However, this is both computationally expensive and intuitively unnecessary because the tracking procedure that relies in short-term (previous frame) information already represents a valid solution for many cases. Therefore, the multiview formulation conceptually decomposes the problem into three tasks: 1) partition processing using short-term information; 2) detection of uncertain regions; and 3) processing of the uncertain regions using multiple views.

Consider an interval of highest posterior probability for e^{t-1} with parameter T_{conf} . Such an interval contains $100 \times T_{\text{conf}}\%$ of the area under the probability density function (pdf) $p(e^{t-1})$. Let $e_{T_{\text{conf}}}^{t-1}$ denote the value of normalized matching error for which $\Pr(e^{t-1} \leq e_{T_{\text{conf}}}^{t-1}) = T_{\text{conf}}$. We can use this parameter as a way of modeling the transition between the single-view and the multiview extraction schemes. On one hand, the short-term

approach is optimal for all those regions whose matching error is reliable. We can express this model as

$$\Pr(\text{svo}_j | \text{noa}^{t-1}, \text{noa}^{t_v}, e^{t-1} < e_{T_{\text{conf}}}^{t-1}, e^{t_v}) = \Pr(\text{svo}_j | \text{noa}^{t-1}) \quad (6)$$

which assumes that the classification is independent of multiview information as long as the matching error in the previous frame remains in the probability interval. In this case, the model reduces to the single-view case. On the other hand, all those regions whose matching error is outside the interval are considered as uncertain regions and should be processed using multiple views in order to further reduce the possibility of misclassification. By defining

$$\text{noa}^{\min} = \begin{cases} \text{noa}^{t-1}, & \text{if } e^{t-1} < e^{t_v} \\ \text{noa}^{t_v}, & \text{otherwise} \end{cases} \quad (7)$$

the corresponding a posteriori probability can be modeled as

$$\Pr(\text{svo}_j | \text{noa}^{t-1}, \text{noa}^{t_v}, e^{t-1} \geq e_{T_{\text{conf}}}^{t-1}, e^{t_v}) = \Pr(\text{svo}_j | \text{noa}^{\min}) \quad (8)$$

such that the uncertain regions will be assigned to either the j th SVO or to its complement using a similar Bayesian framework [14], [8], but in which information from multiple frames is used to find the minimum normalized matching error.

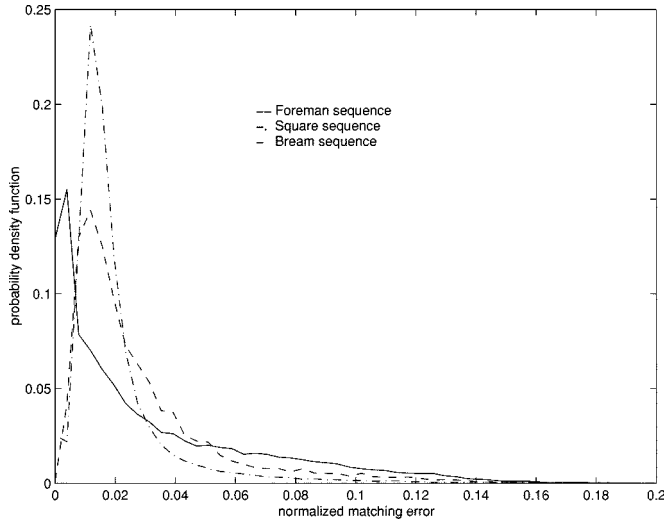


Fig. 4. Approximated pdfs of the normalized matching error $p(e)$ for *Foreman*, *Bream*, and *Square* test sequences. 100 frames were used in each case to compute the pdf.

The application of the MAP criterion to both (6) and (8) is identical: choose H_0 (assign the region R_i^t to the j th SVO) when its probability of belonging to such object so indicates, and choose H_1 otherwise

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

where the variable x takes the values $t-1$ and \min in (6) and (8), respectively. By applying Bayes' rule and rearranging terms

$$\frac{p(noa^x | svo_j)}{p(noa^x | svo_j^c)} \underset{H_1}{\overset{H_0}{>}} \frac{\Pr(svo_j^c)}{\Pr(svo_j)}$$

where the two terms on the left side of the equation represent the two class-conditional pdfs. As shown in [9], these pdfs can be modeled by exponential distributions

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

where $u(\cdot)$ designates the step function. If we let $k^t = \Pr(svo_j^c) / \Pr(svo_j)$, the decision problem can be reduced to the computation of an optimal threshold

$$noa \underset{H_1}{\overset{H_0}{>}} \frac{\lambda_2 - \ln(\lambda_2/k^t \lambda_1)}{\lambda_1 + \lambda_2} = T_{noa}. \quad (10)$$

The threshold can be either computed from the parameters of the model (which in turn can be estimated from data using the last estimated SVO partition), or approximated by 0.5, assuming symmetry between the exponential likelihood functions and $\lambda_i \gg k^t$. Such an approximation has been justified in [9]. A closed expression for the partition operator that extracts the j th SVO from the scene can be written down as

$$P_{SVO_j}^t = \psi_j(P^t) = \{SVO_j^t, E \setminus SVO_j^t\} \quad (11)$$

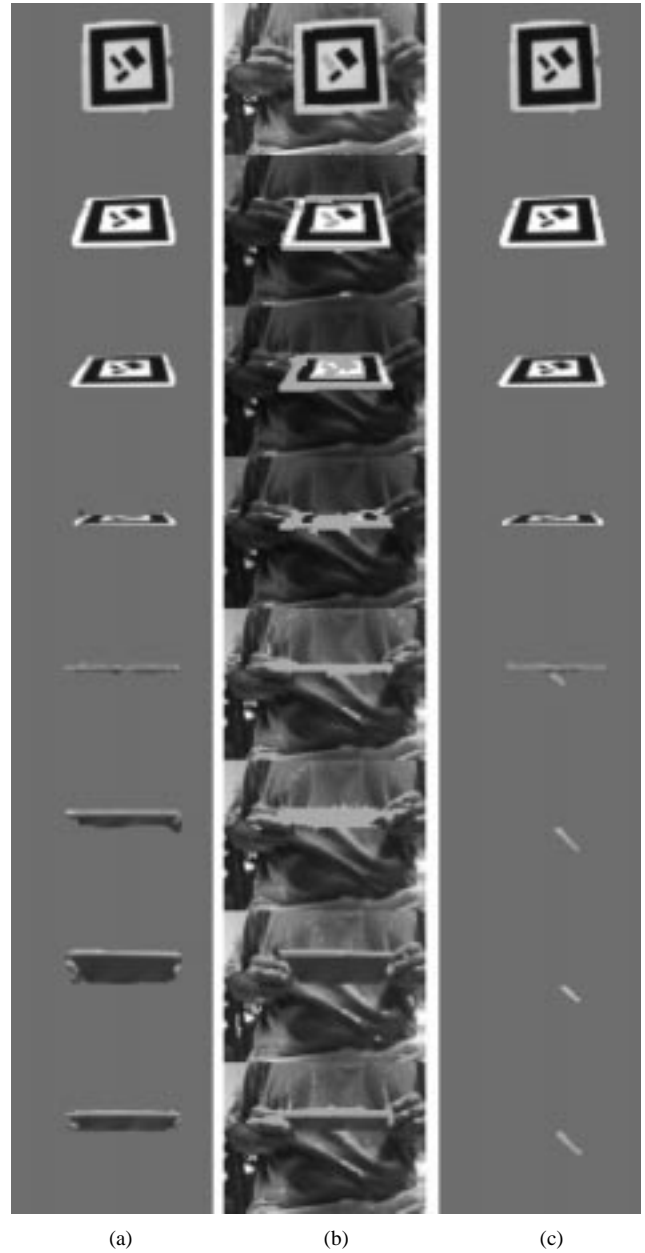


Fig. 5. SVO extraction. *Square* sequence, frames 0, 40, 48, 54, 62, 66, 74, and 82. (a) All regions in the 10% tail of the approximated pdf of the normalized matching error (computed between frames t and $t-1$) are labeled as uncertain. Noisy and disoccluded regions are correctly labeled. (b) SVO extraction using the single-view partition operator. The object is lost when it flips. (c) SVO extraction using the multiview partition operator, for two user-defined views at frames 0 and 68. The object is correctly tracked during the whole sequence.

where $A \setminus B$ denotes set difference, and each SVO has two components

$$SVO_j^t = SVO_{j_{\text{single}}}^t \cup SVO_{j_{\text{multi}}}^t$$

where

$$SVO_{j_{\text{single}}}^t = \bigcup_i R_i^t, \quad noa_{ij}^{t-1} \geq T_{noa}, \quad e^{t-1} < e_{T_{\text{conf}}}^{t-1}$$

and

$$SVO_{j_{\text{multi}}}^t = \bigcup_i R_i^t, \quad noa_{ij}^{\min} \geq T_{noa}, \quad e^{t-1} \geq e_{T_{\text{conf}}}^{t-1}.$$

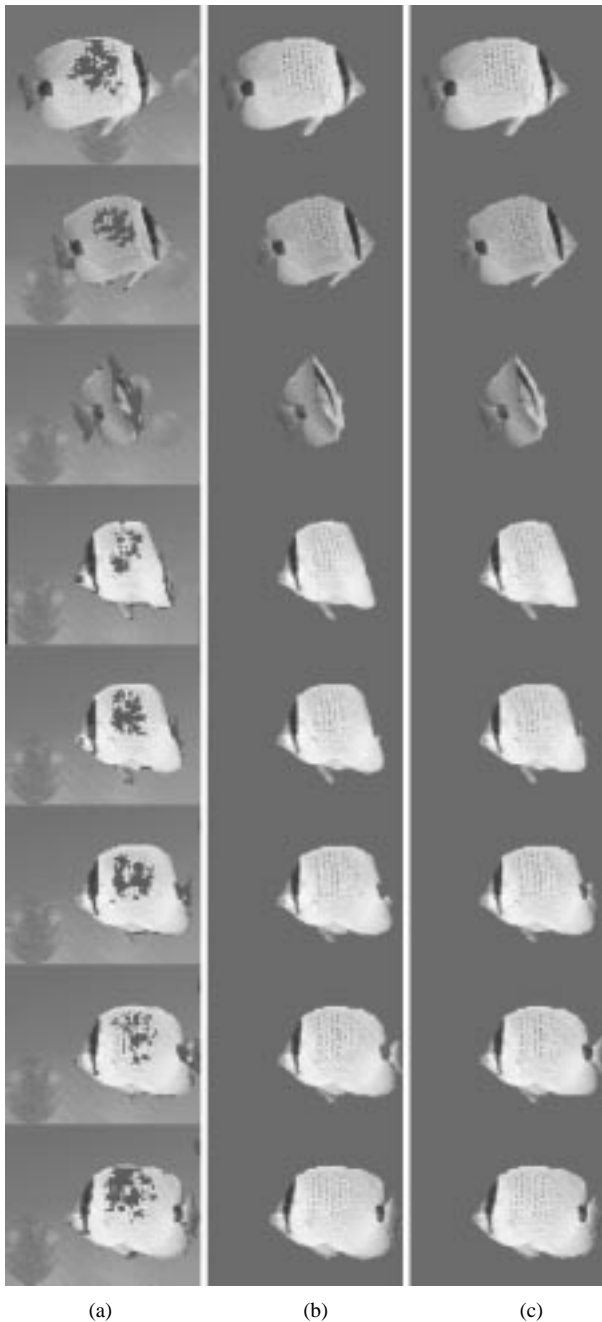


Fig. 6. SVO extraction. *Bream* sequence, frames 0, 100, 114, 122, 124, 126, 128, and 132. (a) Uncertain regions. Textured and disoccluded regions are correctly labeled. (b) SVO extraction using the single-view partition operator. Some detail of the fish tail is lost between frames 123 and 130 due to disocclusion. After that, the tail is recovered. (c) SVO extraction using the multiview partition operator, for two user-defined views at frames 0 and 140. The object is correctly tracked during the whole sequence. The tail is never lost.

By applying the procedure to the M SVOs of the scene, the desired SVO partition P_{SVO}^t can be generated from the set of individual SVO partitions. We define the multiframe SVO extraction operator by

$$P_{SVO}^t = \psi_{MV}(P^t) = \bigwedge_{j=1}^{M-1} \psi_j(P^t) = \bigwedge_{j=1}^{M-1} P_{SVO_j}^t \quad (12)$$

where ψ_{MV} denotes the multiview operator, and \bigwedge represents the infimum of all the single SVO partitions, which consists on

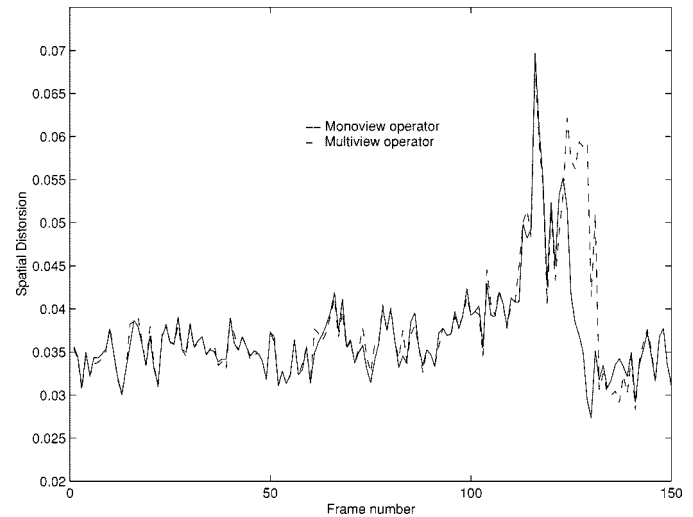


Fig. 7. Spatial distortion for the *Bream* sequence with the single-view (dashed line) and multiview (continuous line) partition operators. The time interval in which the multiview case outperforms the single-view case occurs when the tail of the fish is lost. In the rest of the sequence, the performance is of the same quality.

the partition that is generated by intersecting all the regions in the set $\{P_{SVO_j}^t\}$ (Appendix, Definition 4). Fig. 3 illustrates this operation.

Several theoretical properties of this operator can now be obtained. In fact, it can be shown that it is not order-preserving, and not invertible, and therefore it does not constitute a morphological filter [13].

The generalization of this approach to more than two views is direct. Given K reference SVO partitions ($K > 2$), a set of normalized matching errors $\{e^{tk}, k = 1, \dots, K\}$, and overlapped areas $\{\text{noa}^{tk}, k = 1, \dots, K\}$ can be computed for each region. If we define

$$\text{noa}^{\min} \triangleq \text{noa}^{t^i} \text{ such that } e^{t^i} = \min\{e^{tk}, k = 1, \dots, K\} \quad (13)$$

then the expression for the conditional probability

$$\Pr(\text{svo}_j | \text{noa}^{t^1}, \dots, \text{noa}^{t^K}, e^{t^1}, \dots, e^{t^K})$$

can be reduced to (6) and (8), with the new definition in (13).

The only parameter that remains to be determined is the parameter T_{conf} . Fig. 4 shows the approximated probability density function of the matching error $p(e)$ for several video sequences with very different characteristics. It can be seen that the distributions are concentrated in the low values. We decided to set T_{conf} to a fixed value of 0.9, which is commonly used in statistics as parameter in posterior intervals [14]. Using this value, the value of $e_{T_{\text{conf}}}^{t-1}$ can be computed based on the empirical pdf estimated at each instant. That means that all the regions whose matching error is in the 10% tail of the pdf will be labeled as uncertain regions and reclassified using the other views.

B. Selection of the Reference SVO Partition Set

The creation of the reference SVO partition set is a decision made by the user. It is intended to reflect the different appearances of the objects of the scene (including the background). Recognition of different appearances is a trivial task



Fig. 8. SVO extraction. *Hall* sequence, frames 0, 10, 20, 40, and 50. (a) SVO extraction using the single-view partition operator. (b) SVO extraction using the multiview partition operator, for user-defined partitions at frames 0 and 60.

for a person, provided that he/she can parse the video content in an efficient way. To select the reference SVO partition set, the user can make use of a set of VCR-like capabilities that allow for fast forward and backward access to the video content. Additionally, note that a single view of an SVO might be enough information in many cases (for example, in a head-and-shoulders scene). Therefore, in our implementation, the user can freely specify one or more partitions in the reference set, which would determine the mode of operation (single or multiple views), and also switch between the two modes of operation at any time.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The multiview partition operator was integrated into an SVO extraction system that consists in four stages.

- 1) SVO structure definition by user introduction of the reference SVO partition set (specification of the contours of the objects of interest in K different views).
- 2) SVO computation by spatial partition generation, based on a four-band watershed at each instant [9].
- 3) SVO tracking by application of the multiview partition operator described in the previous section.
- 4) SVO postprocessing for partition refinement.

To test the multiview approach, a two-view reference SVO partition set was selected by the user ($K = 2$). In Fig. 5, we show the results that are obtained with the *Square* sequence presented in Fig. 1 (the test sequences are available in <http://www.hitl.washington.edu/people/danielgp>). Fig. 5(a) illustrates (in light-gray tones) the regions that have been labeled as uncertain, as it was described in Section III. Noisy and disoccluded regions have been correctly labeled. Fig. 5(b) shows the extraction results that are obtained when the single-view partition operator is applied. We observe that the extraction is accurate as long as the object does not completely change its appearance. However, when the other facet of the object appears, there is no way of identifying that some of the new regions in the current partition belong to the same object, and the tracking fails. In comparison, in Fig. 5(c) we show the extraction results obtained with the multiview operator, in which the user selected the two views shown in Fig. 1 as the reference SVO partition set. The extraction quality is equivalent for the first part of the sequence, but it considerably improves when the object flips, as the operator correctly assigned the uncertain regions to the *Square* SVO, allowing for a more accurate extraction. Note that even though a perfect (manual) partition was available for each keyframe of the SVO reference partition set, no partition substitution (reset) was performed during the tracking process. In other words, after the definition of the reference set, the tracking results reported in this section only use automatically generated partitions. Substitution could obviously be done and would improve the results (reducing the spatial distortion to zero for some frames), but we were interested in evaluating the performance of the system without such substitution.

Another example, for which an objective evaluation is also possible, is shown in Fig. 6. Fig. 6(a) illustrates the detected uncertain regions (in dark gray) for a series of frames of the *Bream* sequence. We observe that these regions correspond to textured regions (part of the body of the fish) and to disoccluded regions. The comparison of the extracted SVOs using the single-view and multiview operators is presented in Fig. 6(b) and (c), respectively. On one hand, the extraction is of good quality in the single-view case, with the exception of a few frames for which part of the tail of the fish is temporally lost due to sudden disocclusion, but later recovered. On the other hand, the multiview operator, for which frame numbers 0 and 140 were selected as the reference set, recovers the fish without losing track of any part of it. The objective evaluation, based on the spatial distortion measurement proposed in [24], is presented in Fig. 7. We observe that the multiview operator improves the performance when the single-view operator fails, without degrading the performance at any time.

To further illustrate the performance of our methodology, two experiments were performed for the difficult *Hall* test sequence

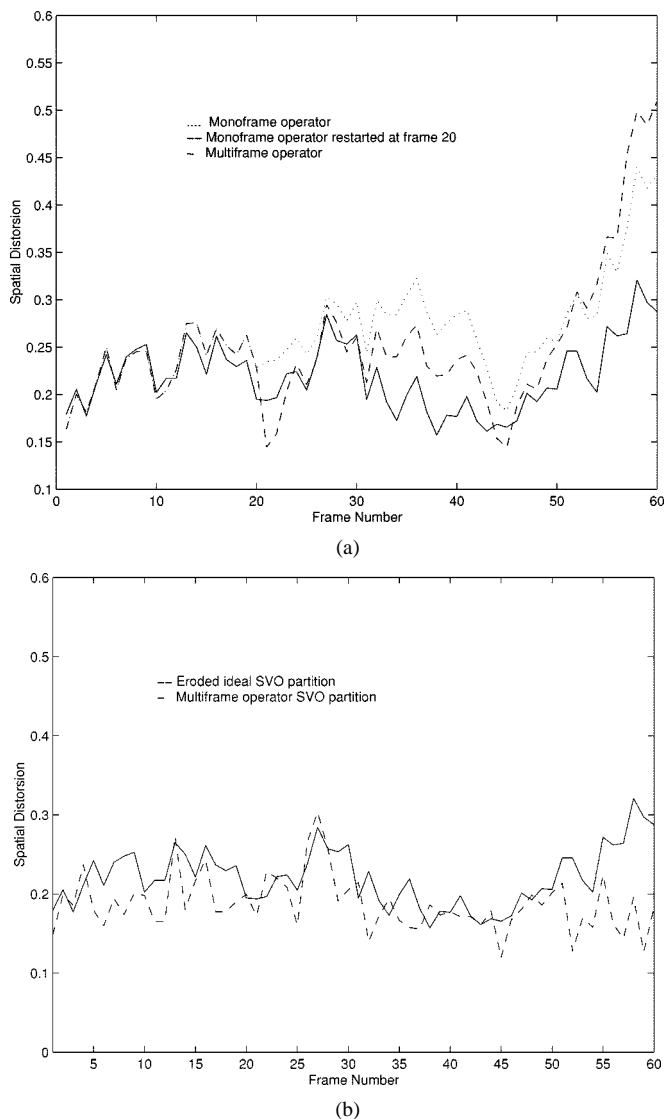


Fig. 9. (a) Spatial distortion for the *Hall* sequence with the single-view operator (dotted line), single-view operator restarted at frame 20 (dashed line), and multiview operator (continuous line). The multiview case outperforms the other two cases. (b) Comparison between the distortion of the multiview operator (continuous line) compared to the distortion of a 3×3 -eroded version of the ideal (manually generated) partition, which peels off the hand-made partition by one pixel (dashed line).

that has been partitioned into the Man SVO and the rest of the scene. In the first place, the single-view operator is applied to 60 frames, with the system starting at frame number 0. In the second place, our proposed two-view operator, with manual definition of SVO partitions at frames number 0 and 60, is also applied. Fig. 8(a) and (b) shows some of the results obtained with the two operators. Additionally, Fig. 9(a) shows the corresponding spatial distortion. We observe that, even though the two operators behave similarly for the first frames of the sequence, the multiview operator produces more stable and accurate masks in the long term. In particular, note that the multiview frame can handle the case of separation of the Man SVO into two parts (man and suitcase) and tracks both parts with accuracy. Furthermore, the use of a multiview operator reduces the amount of user interaction with respect to the number of manual SVO partitions that are needed. To illustrate this case, a third exper-

iment was performed. The user stops the single-view tracking process at $t = 20$ (frame number in which the single-view operator performance starts degrading compared to the multiview case), manually generates the corresponding SVO partition, and restarts the single-view system for the rest of the sequence. We can observe from Fig. 9(a) that the multiview operator performs better once again. As opposed to the single-view case, in which user interaction only affects the short term for tracking purposes, the multiview operator integrates the set of views and allows for a better decision on disoccluded regions. Finally, to provide an idea of the degree of accuracy of the generated SVO partitions, we compare the spatial distortion obtained with the multiview operator and the distortion computed between the ideal partition and a 3×3 -eroded version of itself in Fig. 9(b). The erosion operation has peeled off the hand-made partition by one pixel. We observe that both distortion figures are comparable.

We have conducted other experiments to analyze the dependence of our method with respect to the required parameters. In particular, there are two parameters that need to be set: the parameter of the interval of highest posterior probability T_{conf} and the search window size w for the estimation of the regional motion vectors (necessary for the computation of the normalized overlapped area for each region). On one hand, the first parameter determines the amount of regions for which multiple view information will be used for region classification. As we discussed in Section III, for many cases, single-view information is enough to correctly assign each region to the correct SVO. Therefore, a high value for T_{conf} should produce good quality results while keeping reasonable computations. On the other hand, the search window establishes the domain on which the motion vector can be estimated. As reported in [12], a fixed size of w works for backward tracking (a single-view approach) in several different scenarios. However, when the multiview operator is used, it could be useful to set a larger search window w_u for the multiview matching procedure, such that a better regional correspondence can be found for the uncertain regions.

In Fig. 10(a), we show the curves of spatial distortion as a function of T_{conf} , for the *Hall* sequence. It can be seen that the distortion figures generated for $T_{\text{conf}} = 0.7, 0.8,$ and 0.9 are very similar to each other, which suggests that the selection of this parameter is not so crucial. The same behavior (both quantitatively and qualitatively) has been observed for other test sequences. In Fig. 10(b), we present the results of the spatial distortion as a function of the uncertain region search window size w_u (while keeping the single-view window size fixed to $w = 10$). The results for $w_u = 10$ and $w_u = 15$ are virtually identical (only the latter is shown). In fact, the window size introduces a well-known tradeoff: larger searches may improve tracking of larger motion, but if the scene is complex, the possibility of false regional matching also increases. For *Hall*, in which the scene background is highly cluttered, the augmentation of the window size to $w_u = 20$ and $w_u = 25$ does not introduce larger distortion in the first frames of the sequence. But as time evolves, the false matchings tend to degrade the extraction performance. Multiscale methods might offer good solutions to some of these situations.

An example that illustrates the advantages of the multiview operator for tracking of multiple objects can be appreciated in

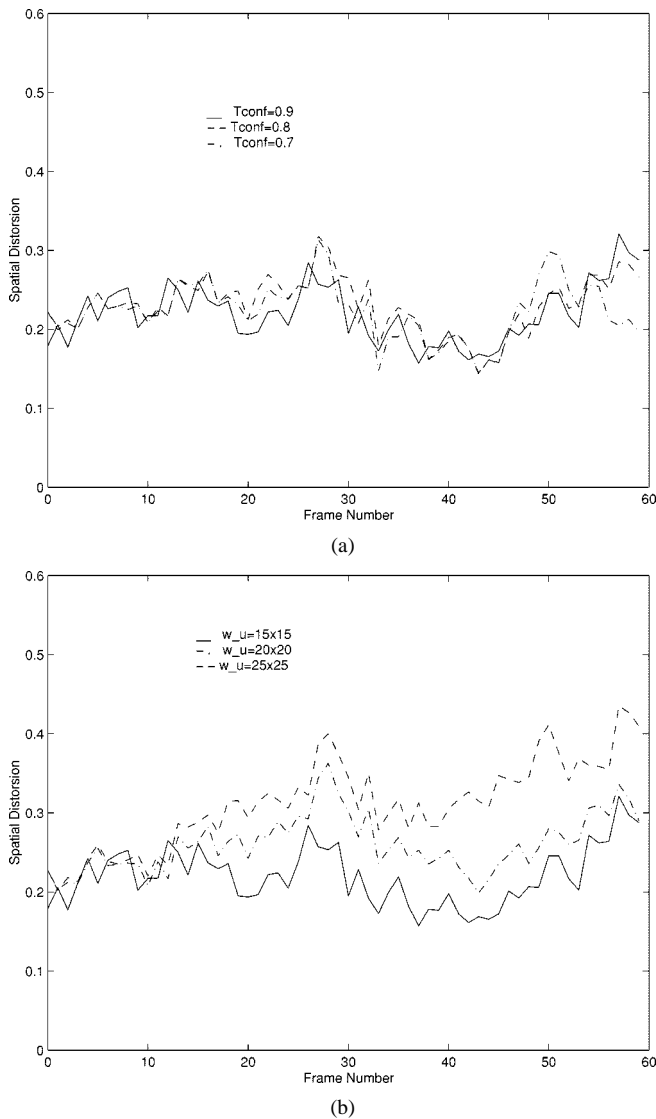


Fig. 10. (a) Spatial distortion produced by the multiview operator as a function of the parameter of the interval of highest probability, T_{conf} . *Hall* sequence. Three values, equal to 0.7 (dotted-dashed line), 0.8 (dashed line), and 0.9 (continuous line) are shown. An approximately similar spatial distortion is produced in all cases. (b) Spatial distortion as a function of the uncertain region search window size w_u : 15×15 (continuous line), 20×20 (dotted-dashed line), and 25×25 (dashed line). The search window for the short-term tracking was fixed to $w = 10$. An increasing window size for uncertain regions will allow for tracking of larger displacements, but is also susceptible to motion estimation errors (false matchings), which increases the spatial distortion. The results in this paper use $w_u = 15$.

Fig. 11 for the *Apple* sequence. In this sequence, the apple and the hand originally have the same motion, but when the hand lets the apple roll free in the table, multiple object occlusions/disocclusions (hand, apple and background) occur. The result of applying the single-view operator is shown in Fig. 11(a) and (b). We observe that when the objects split, the disoccluded region is not correctly assigned to the background. The results of applying the multiview operator, using frames 0 and 90 for the reference SVO partition set, is shown in Fig. 11(c) and (d). In this case, when a view of the background is provided, the method correctly detects the disocclusion regions and allows the splitting and remerging of the objects as time evolves. Note that, due to the random motion of the apple, a user would need to make

correction of the results at several frames (i.e., each time the background is covered and then uncovered) if a single-view approach were chosen, to get the quality that is obtained with the multiple view technique. As this situation occurs frequently in real-world sequences, a multiple-view approach offers advantages in terms of amount of user interaction.

One important issue is the sensitiveness of the performance with respect to the selection of the reference SVO partition set. The performance of the method obviously depends of the views that are selected as input by the user. However, the specific selection of the keyframes is not critical as long as the chosen frames approximately describe the same content. An example of this point performed with the *Apple* sequence is shown in Fig. 12. Fig. 12(a) displays the spatial distortion introduced for three cases of the two-view operator, where frame 0 is selected as the first view, and frames 74, 84, and 117 are selected as the second view, respectively. Such frames contain similar content. We can observe that the spatial distortion is practically identical in all cases before the objects split, and remains somewhat similar in the rest of the sequence. Fig. 12(b)–(d) illustrate each extraction process at frames 80, 90, and 110, and show that there occurred some errors in the extraction process for two of the three cases around frame 90. However, in all cases, the system subsequently extracts the two objects with good quality.

Another example that illustrates tracking of multiple objects in moderate clutter is presented in Fig. 13, for the *Handshake* sequence [16]. In this case, the heads of two men who pass in front of each other are tracked. Fig. 13(a) presents the results for the single-view operator. We can observe that extraction errors occur when the man on the left disoccludes parts of the background (frames 25 and subsequent), and again when the man on the right walks by the other man and also disoccludes the background (frames 50 and subsequent). On the contrary, when two views are selected (frames 0 and 34), the tracking of the two heads is accurate along the whole sequence Fig. 13(b). In addition, and similarly to the previous example, the selection of the second view is not critical, as long as the uncovered background of the scene is exposed in such view. Furthermore, more than two views provide a redundant but more robust SVO representation.

We are currently integrating our extraction algorithms in a general interactive system for analysis of digital video. One example of SVO extraction in consumer video materials appears in Fig. 14 for the *Cat* sequence, which has hand-held camera motion, zooming, and independent motion of the cat. We have tested our algorithms with other sequences that present object disocclusions including object disappearance/reappearance, obtaining similar extraction results.

The computational complexity is adequate for semiautomatic SVO extraction. The extraction takes about 3.5 s/frame in QCIF color images, on an SGI Octane computer. This figure could be significantly reduced by code optimization and several simplifying operations (like computing segmentation and tracking only on subregions of the image).

The limitations of our approach arise when extracting SVOs in cluttered scenes where the colors of different SVO are very similar, and in videos of poor quality, as spatial segmentation is likely to contain errors in both situations. Furthermore, in cases



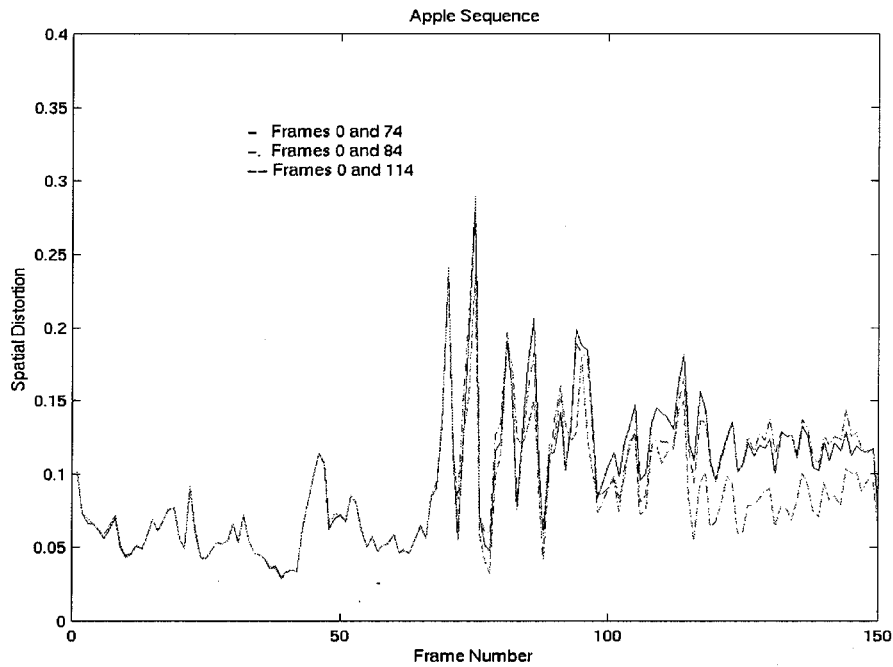
Fig. 11. Multiple SVO extraction. *Apple* sequence, frames 0, 10, 40, 50, 62, 72, 76, 80, 84, 88, 92, 100, 112, 120, 130, 150. (a)–(b) Tracking using the single-view partition operator. Due to sudden background disocclusion, regions of the background are erroneously assigned to the SVOs. (c)–(d) Tracking with the multiview partition operator, in which frames 0 and 90 were used as reference partition set. The objects are correctly tracked.

like indoor material recorded with nonprofessional quality, the natural colors of the scene are usually affected by interior light and the color camera response. This situation also hardens segmentation. Additionally, the spatio-temporal features become less reliable as clutter increases. In these cases, both segmentation and classification errors will likely be introduced in the SVO partitions and propagated along time. This constitutes a

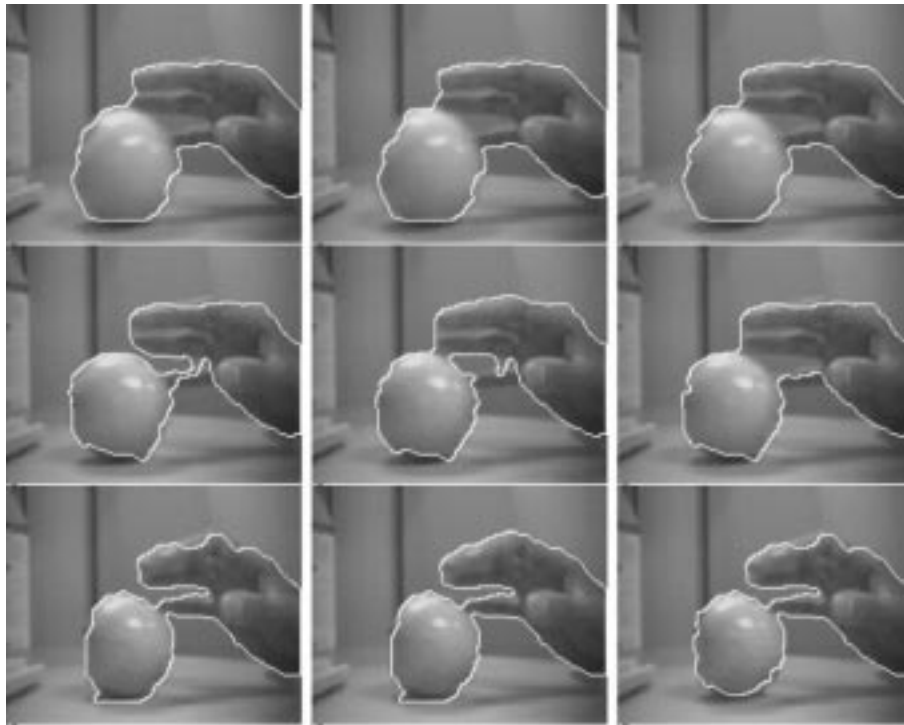
common limitation to all segmentation-based SVO extraction methods.

V. CONCLUSION

In this paper, we present a multiview morphological approach for SVO extraction. Our methodology is based on the design



(a)



(b)

(c)

(d)

Fig. 12. SVO extraction. *Apple* sequence, selection of the reference partition set. (a) Spatial distortion introduced by the two-view extraction operator, where frame 0 is fixed for all cases, and the second view is different in each case: frame 74 (continuous line), frame 84 (dashed-dotted line), and frame 114 (dashed line). The quality in all cases remains approximately the same. SVO extraction at frames 80, 90, and 110 using: (b) frames 0 and 74; (c) frames 0 and 84; and (d) frames 0 and 117. Note that the error present at frame 90 (finger not extracted) is later solved by all the cases (frame 110 and subsequent).

of extensive operators on lattices of partitions that apply a regional MAP principle for region classification, and addressed an issue of practical importance: how to reduce the tracking problems related with SVO disocclusions in off-line applications. We have shown that this approach improves the SVO extraction quality compared to single-view schemes because in-

formation from multiple scenes can solve for cases in which single-view methods fail due to object disocclusions. The probabilistic approach allows for the formulation of these operators using optimality criteria, while their algorithmic complexity remains low. From this perspective, our system for SVO extraction is not the merging of disparate algorithms, but the application



Fig. 13. SVO extraction. *Handshake* sequence. The heads of the two men are tracked. (a)–(b) Single-view SVO extraction. (a) Frames 0, 10, 20, 25, and 30; (b) Frames 35, 40, 45, 50, and 55. (c)–(d) Two-view SVO extraction (frames 0 and 34 are used as reference SVO partition set). Same frame numbers.

of a common methodology. From our investigation, we are currently interested in three main lines of research: 1) the definition of other partition operators and the study of their algebraic properties; 2) the study of methods for better probabilistic modeling to increase their robustness; and 3) the development of improved criteria for objective evaluation of SVO extraction.

APPENDIX

The first three definitions and the last one are adapted from [22] and [13]. The rest are extracted from [9].

Definition 1: Given a space E and its collection of subsets $\mathcal{P}(E)$, a *partition* P of E is a mapping $P: E \rightarrow \mathcal{P}(E)$, such that $\forall x, y \in E$, the two following conditions hold: 1) $x \in P(x)$ and 2) $P(x) = P(y)$ or $P(x) \cap P(y) = \emptyset$, where \emptyset denotes the empty set. $P(x)$ is called the *zone* or *region* of P that contains x . 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 $R_i = \cup x \in E$ such that $P(x) = R_i$.

Definition 2: A *complete lattice of partitions* Π is a set of partitions of E with a partial ordering relation (\leq), for which each of its subsets has an infimum (\wedge) and a supremum (\vee).

Definition 3: *Partial ordering* relation between partitions. Let P_i and P_j be two partitions of E . P_i is said to be *finer*

(smaller) than P_j iff $P_i(x) \subseteq P_j(x), \forall x \in E$, and it is represented by $P_i \leq P_j$.

Definition 4: The *infimum* of a set of partitions indexed by i is defined by

$$\left(\bigwedge_i P_i \right) (x) = \cap_i P_i(x) \quad \forall x \in E$$

i.e., it corresponds to the partition made of the intersections of all the regions in the original set of partitions. Additionally, the *supremum* is defined by

$$\left(\bigvee_i P_i \right) (x) = \cap \{B: B = \cup_i \cup_{y \in B} P_i(y), x \in B, B \in \mathcal{P}(E)\}$$

which is the finest partition that is larger than each of the individual partitions. 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 5: 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 $P_I(x) = E$ for all $x \in E$.

Definition 6: Let $\mathbf{I} = \{\mathbf{I}^t | t \in \mathcal{Z}\}$ be a multivalued image sequence, with domain $E = \mathcal{D}(\mathbf{I}^t) \subset \mathcal{Z}^2$. Let $P^t = \{R_i^t, i \in \{1, \dots, N\}\}$ denote a partition of E at time t . The j th *semantic*

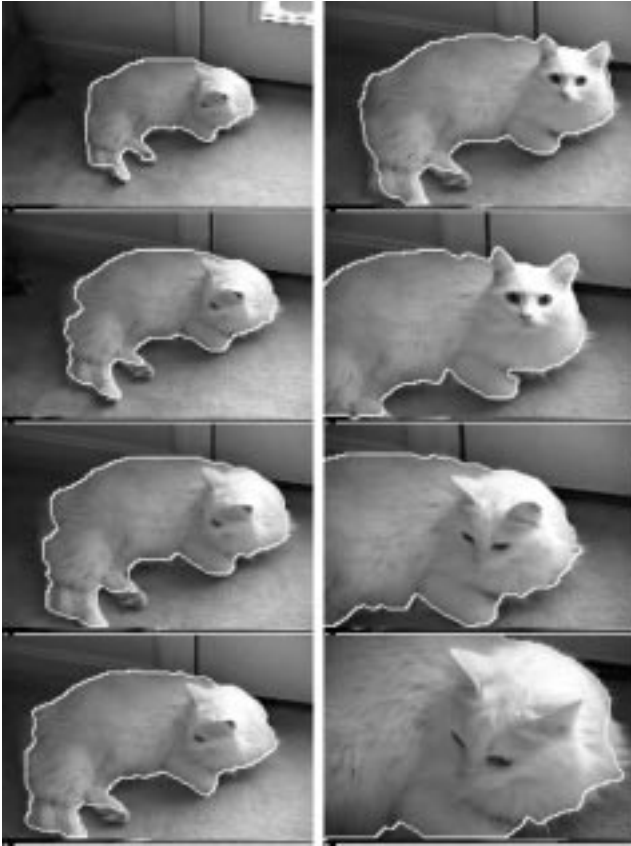


Fig. 14. SVO extraction. *Cat* Sequence. Frames 0 and 500 were used as reference SVO partition set. (a) Frames 0, 30, 50, and 60. (b) Frames 350, 430, 450, and 500.

video object of the scene depicted in \mathbf{I} (consisting of M objects) is defined by $\text{SVO}_j = \{\text{SVO}_j^t\}$, where

$$\text{SVO}_j^t = \bigcup_{i=1}^{N_j^t} R_i^t$$

and where SVO_j^t represents the j th VOP at time t , each composed of N_j^t regions of P^t .

Definition 7: An SVO partition P_{SVO}^t is the collection $P_{\text{SVO}}^t = \{\text{SVO}_j^t, j \in \{1, \dots, M\}\}$, where M is the current number of SVOs in the scene.

Definition 8: Let Π be a complete lattice of partitions of $E = \mathcal{D}(\mathbf{I}^t)$. Let $P^t \in \Pi$, and let $TR = \{P_{\text{SVO}}^{t_k}, k = 1, \dots, K\}$ denote a reference SVO partition set, composed of K SVO partitions of the scene at time instants t_k , i.e., partitions that correspond to different scene views. An SVO extractor operator is a mapping $\psi: \Pi \rightarrow \Pi$, such that $P_{\text{SVO}}^t = \psi_{TR}(P^t)$. As the set TR is used as a reference, its explicit inclusion in the notation is usually omitted.

Definition 9: The normalized overlapped area between a region R_i^t of a partition at time t and the j th SVO at time t_k is defined by

$$\text{noa}_{ij}^{t_k} = \frac{\text{card}([R_i^t]_{V_i^{t_k}} \cap \text{SVO}_j^{t_k})}{\text{card}(R_i^t)}$$

where $V_i^{t_k}$ denotes the motion vector estimated for R_i^t between frames at times t and t_k , assuming a translational motion model, $[R_i^t]_{V_i^{t_k}}$ represents the motion-compensated version of R_i^t , and $\text{card}(A)$ denotes the cardinality of the set A .

Definition 10: The normalized matching error between a region R_i^t of a partition at time t , and the image sequence at time t_k is defined by

$$e_i^{t_k} = \frac{\sum_{x \in [R_i^t]_{V_i^{t_k}}} |\mathbf{I}^t - \mathbf{I}^{t_k}|}{255 \cdot \text{card}(R_i^t)}$$

Definition 11: A partition operator ψ is called extensive iff $P \leq \psi(P), \forall P \in \Pi$.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments, which improved the content of the paper. They also thank Dr. J. MacCormick for providing the *Handshake* sequence.

REFERENCES

- [1] A. Baumberg and D. Hogg, "Learning deformable models for tracking the human body," in *Motion Based Recognition*, M. Shah and R. Jain, Eds. Norwell, MA: Kluwer, 1997, pp. 39–60.
- [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*. Berlin, Germany: Springer-Verlag, 1988.
- [4] R. Castagno, T. Ebrahimi, and M. Kunt, "Video segmentation based on multiple features for interactive multimedia applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, Sept. 1998.
- [5] P. Correia and F. Pereira, "The role of analysis in content-based video coding and indexing," *Signal Processing*, vol. 66, no. 2, pp. 125–142, Apr. 1998.
- [6] E. Chalom and V. M. Bove, "Segmentation of an image sequence using multi-dimensional image attributes," in *Proc. ICIP'96*, vol. 2, Lausanne, Switzerland, 1996, pp. 525–528.
- [7] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*. Cambridge, U.K.: Cambridge Mathematical Textbooks, 1990.
- [8] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2000.
- [9] D. Gatica-Perez, C. Gu, and M. T. Sun, "Semantic video object extraction using four-band watershed and partition lattice operators," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 603–618, May 2001, submitted for publication.
- [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 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.
- [12] —, "Semantic video object tracking using region-based classification," in *Proc. ICIP'98*, Chicago, IL, Oct. 1998, pp. 643–647.
- [13] H. J. A. M. Heijmans, *Morphological Image Operators*. New York: Academic, 1994.
- [14] P. M. Lee, *Bayesian Statistics*. New York: Wiley, 1997.
- [15] H. Luo and A. Eleftheriadis, "Designing an interactive tool for video object segmentation and annotation," in *Proc. ACM Multimedia Conf.*, Orlando, FL, 1999.
- [16] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," in *Proc. 7th Int. Conf. Computer Vision*, vol. 1, Kerkiras, Greece, 1999, pp. 572–578.
- [17] F. Marques and J. Llach, "Tracking of generic objects for video object generation," in *Proc. ICIP*, Chicago, IL, Oct. 1998, pp. 628–632.
- [18] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-D objects from appearance," *Int. J. Comput. Vis.*, vol. 14, pp. 5–24, 1995.

- [19] Microsoft Windows Media Player™ (1999). [Online]. Available: <http://www.microsoft.com/windows/windowsmedia/en/default.asp>
- [20] N. O'Connor and S. Marlow, "Supervised semantic object segmentation and tracking via EM-based estimation of mixture density parameters," in *Proc. Noblesse Workshop on Non-linear Model Base Image Analysis*, Glasgow, Scotland, 1998.
- [21] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1993.
- [22] J. Serra, *Image Analysis and Mathematical Morphology, Vol. II: Theoretical Advances*. New York: Academic, 1988.
- [23] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE-CVPR*, Maui, HI, June 1991, pp. 586–591.
- [24] M. Wollborn and R. Mech, "Refined procedure for objective evaluation of video object generation algorithms," Doc. ISO/IEC JTC1/SC29/WG11 M3448, Mar. 1998.
- [25] D. Zhong and S.-F. Chang, "An integrated approach for content-based video object segmentation and retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 1259–1268, Dec. 1999.



Daniel Gatica-Perez (S'01) received the B.S. degree in electronics 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.



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, Seattle, in August 1996, where he is a Professor. Previously, he was the Director of the Video Signal Processing Research Group at 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 the current Editor-in-Chief of IEEE TRANSACTIONS ON MULTIMEDIA and a former Editor-in-Chief of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) from 1995 to 1997. 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 was a conference General Co-Chair of the Visual Communications and Image Processing 2000 Conference. He received an Award of Excellence from Bellcore in 1987 for work on Digital Subscriber Line, was a co-recipient of the TCSVT Best Paper Award in 1993, and received a Golden Jubilee Medal from the IEEE CAS Society in 2000.



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. research was on multivalued morphology and segmentation-based coding and tracking.

From 1989 to 1990, he was an Assistant Director at the CAD center of Fudan University. He was an Associate Research Fellow at the European Organization for High Energy Physics (CERN) in 1991. During 1992 to 1995, he worked at 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. He has also been an Affiliate Professor at the University of Washington since 1999. He has published more than 40 papers and earned six patents on these subjects.

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