Spatiotemporal Segmentation Based on Region Merging

Fabrice Moscheni, *Member, IEEE,* Sushil Bhattacharjee, and Murat Kunt, *Fellow, IEEE*

Abstract—This paper proposes a technique for spatiotemporal segmentation to identify the objects present in the scene represented in a video sequence. This technique processes two consecutive frames at a time. A region-merging approach is used to identify the objects in the scene. Starting from an oversegmentation of the current frame, the objects are formed by iteratively merging regions together. Regions are merged based on their mutual spatiotemporal similarity. The spatiotemporal similarity measure takes both temporal and spatial information into account, the emphasis being on the former. We propose a Modified Kolmogorov-Smirnov test for estimating the temporal similarity. This test efficiently uses temporal information in both the residual distribution and the motion parametric representation. The region-merging process is based on a weighted, directed graph. Two complementary graph-based clustering rules are proposed, namely, the strong rule and the weak rule. These rules take advantage of the natural structures present in the graph. Also, the rules take into account the possible errors and uncertainties reported in the graph. The weak rule is applied after the strong rule. Each rule is applied iteratively, and the graph is updated after each iteration. Experimental results on different types of scenes demonstrate the ability of the proposed technique to automatically partition the scene into its constituent objects.

Index Terms—Automatic spatiotemporal segmentation, object segmentation, region merging, modified Kolmogorov-Smirnov test, weighted directed graph.

1 Introduction

SPATIOTEMPORAL segmentation techniques attempt to identify the objects present in a scene based on spatial and temporal (motion) information [1], [2]. We define *spatial information* as being the brightness information and *temporal information* as being the motion information. The scene is partitioned into regions such that each region (except the background) represents a moving object. According to the Gestalt "law of common fate," meaningful regions are obtained if the regions are defined on the basis of temporal coherence. Consequently, the resulting regions can be identified as moving objects composing the scene [3]. Such a segmentation provides an alternative to the waveform representation of the visual information. In contrast to the latter representation which directly derives from the image capture process, the segmentation in terms of objects describes the content of the scene. Not only is it independent of the image capture process, but it is also semantically meaningful.

Spatiotemporal segmentation plays a fundamental role in computer-assisted scene analysis. Spatiotemporal segmentation forms the backbone of schemes for recognizing and classifying objects or tracking them. Thus, spatiotemporal segmentation has important applications in fields such as robot vision (for identifying and tracking objects) and video coding (for realizing so-called second-generation video-coding approaches [4]). Spatiotemporal segmentation also finds application in creating mosaics, where the goal is to generate time-integrated views of a scene [5]. This is useful in browsing libraries of digital video sequences.

By its very nature, the problem of defining the objects forming a scene is a paradox. There is indeed a strong interdependence between the estimation of the spatial support of an object and of its motion characteristics. On one hand, estimation of the motion characteristics of the object depends on the region of support of the object. Therefore, an accurate segmentation of the object is needed in order to estimate the motion accurately. On the other hand, a moving object is characterized by coherent motion characteristics over its entire region of support (assuming that only rigid motion is permitted). Therefore, an accurate estimate of the motion is required in order to obtain an accurate segmentation of the object. Furthermore, accurate object definition involves not only motion information, but also spatial characteristics. In particular, the spatial information provides important hints about object boundaries. However, the best strategy for combining the two types of information remains an open issue.

Faced with the challenges of spatiotemporal segmentation, some techniques rely not just on the information available in two consecutive frames. One way of increasing the amount of available information is to extend the temporal support used for analysis [6], [7], [8]. The segmentation procedure can thus utilize the information present in more than two frames. Segmentation may be further facilitated by requiring the scene to satisfy a given set of constraints. These constraints may be interpreted as a priori knowledge which is incorporated in the segmentation procedure. For instance, the number of objects present

- *The authors are with the Signal Processing Laboratory, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland.*
  E-mail: fabrice.moscheni@ epfl.ch, sushil.bhattacharjee@epfl.ch, murat.kunt@epfl.ch.

Manuscript received 6 May 1997; revised 8 July 1998. Recommended for acceptance by D.J. Kriegman.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 107136.

in the scene may be predefined [9], [10]. However, such approaches are not general. Either the segmentation process can no longer be performed automatically or it addresses only certain types of scenes. This loss of generality may be unacceptable in applications such as video coding.

In this paper, we present a region-merging technique for spatiotemporal segmentation that relies only on the information existing in two consecutive frames of a video sequence. No a priori knowledge is assumed about the number or characteristics of objects in the scene. Starting from an oversegmentation of the current frame, the proposed technique determines the objects constituting the scene at hand. To that end, the regions in the oversegmentation are merged according to their mutual spatiotemporal similarity. We propose a novel measure to assess the spatiotemporal similarity between regions. This measure combines temporal and spatial information. The information about spatiotemporal similarity between regions is represented in the form of a graph. A graph-based hierarchical clustering algorithm is used to detect clusters of similar regions, based on two clustering rules proposed here.

The paper is laid out as follows. In Section 2, we present a discussion of the techniques for spatiotemporal segmentation that have been proposed in the recent past. Their relative merits and shortcomings are discussed. Section 3 gives an overview of the proposed technique. The important components of the proposed algorithm are discussed in detail in the following sections. The proposed region-merging technique takes as input an oversegmentation of the current frame. One method for generating such sets of initial regions is described in Section 4. The spatiotemporal similarity measure used for merging regions is described in Section 5. The similarity information among the regions is represented in a graph. The graph-based clustering strategy used to detect clusters of similar regions is presented in Section 6. Section 7 presents experimental results produced by the proposed method. Concluding remarks are presented in Section 8.

2 Previous Work

Techniques for spatiotemporal segmentation can generally be grouped into two categories. Some techniques take a top-down approach as they sequentially extract the different objects forming the scene. Other techniques have a bottom-up approach. These typically start with an oversegmentation of the image and iteratively merge regions in order to arrive at a coherent and stable description of the scene.

The top-down methods rely on the outlier detection/rejection paradigm. The objects are sequentially extracted by iteratively determining the successive dominant motion characteristics [11], [12], [13], [14]. Pixels complying with the current dominant motion are assumed to comprise one object. The other pixels are seen as outliers. Only these outlier pixels are considered in the next iteration for estimating the subsequent dominant motion and the corresponding objects. The top-down approach is faced with two major problems. The first problem is the estimation of the dominant motion characteristics in the presence of multiple local motion characteristics. This dilemma is known as the generalized aperture problem [10]. The second problem arises from the fact that outliers influence the estimation of the motion characteristics.

Techniques based on pyramidal decomposition of the image have been proposed to overcome the first problem [15]. Such a decomposition separates the different motions existing in the scene. In the presence of a dominant motion, the estimation procedure locks on to this motion and ignores the others. The pyramidal approach has been used by the top-down techniques proposed in [9], [11], [16], [17]. The results can be further improved by the use of robust estimators that can identify the dominant motion without being influenced by the other local motions existing in the scene [18], [19], [20].

Most techniques [17], [19] rely on temporal information to detect outliers. The segmentation of the previous frame is warped onto the current frame, using the estimated parameters of the dominant motion. The outliers are defined as regions corresponding to large prediction errors. However, this information may be unreliable, especially in low-activity regions. The outlier detection procedure can be made more robust by integrating information from the residuals of large regions [21] and spatial information [20], [22], [23].

Top-down approaches are characterized by simplicity and low computational complexity. However, the process of successively determining the characteristics of the remaining dominant motion imposes an artificial hierarchy among the objects in the scene. Furthermore, a complete partitioning of the frame is not guaranteed by top-down methods. The successive extraction of the dominant motions may lead to a situation where the remaining outlier pixels do not belong to any object.

Bottom-up approaches rely on a region-merging procedure to identify meaningful objects. First, a set of initial regions is derived. Usually these regions do not represent meaningful objects. Bottom-up approaches merge these regions, based on some measure of spatiotemporal similarity, so as to obtain meaningful moving objects. Such approaches may be decomposed into three steps: the creation of the set of initial regions, the definition of the region similarity measure, and how this measure is used to merge the regions into objects.

Different approaches proposed in the past have used different techniques for generating the set of initial regions. Some authors simply assume each pixel represents a region [1], [2], [24]. Although such regions are guaranteed to be spatiotemporally coherent, they have no inherent meaning and do not contribute to any new information to the region-merging process. Estimates of motion for individual pixels are very unreliable. Also, this approach is computationally very expensive. A simple quadtree-based segmentation approach has also been used to generate the initial regions [25]. Szeliski and Shum [26] use quadtree splines for motion estimation. The spatial delimitation of such regions often does not reflect the true spatial structures present in the scene. This disadvantage is overcome by using arbitrary-shaped initial regions which are spatiotemporally homogeneous [3], [27].

The region-merging process is based on some spatiotemporal similarity measure. Typically, the similarity measure relies mainly on temporal information [3], [28], [29], [30], [31], [32]. The temporal information is represented in a fully parametric form [28]. The number and characteristics of the parameters depend on the model chosen. Examples of motion models are the translational model (two parameters), the affine model (six parameters), and the perspective model (nine parameters) [33].

When assessing the temporal similarity between regions, the parametric motion representation may be used in several ways. The methods of Dufaux et al. [3] and Wang and Adelson [29] define the region similarity measure to be the distance between the corresponding sets of motion parameters in the motion parameter space. This clustering approach is sensitive to errors in the parametric representation and to the distance measure used in the clustering process. Also, a given optical flow may be described by several different parametric representations [30]. This implies that two regions with similar motion may turn out to have very different motion parameters and, thus, may end up in different clusters. The other way of using the motion information is based on the statistics of the residual distributions obtained after motion compensation of the regions in question. For instance, the similarity measure proposed by Adiv [2] is the variance of the residual distribution. However, the use of a single statistic is too drastic a reduction of information and may lead to wrong merging decisions.

The definition of the region similarity is a challenging issue. All the available information should be put to work in order to robustly define the objects present in the scene. The similarity measure should exploit both spatial and temporal information. However, the best strategy to combine both sources of information remains an open issue. Several approaches have been proposed [2], [34], [35], [36]. Thompson merges regions on the basis of a contrast criterion modified through temporal information [37]. However, the two types of information are not combined into a single similarity measure. Wandell [38] states that visual sensitivity depends jointly on space and time. This nonseparability may be a helpful guideline for defining an efficient measure of region similarity. Nevertheless, as the primary characteristic of an object is its coherent motion, the emphasis should remain on the temporal information.

The region-merging strategy should make the best use of the information provided by the similarity measure. The strategy should be robust against errors or ambiguities and should also be computationally viable. Some region-merging approaches incorporate a two-step strategy. First, the ensemble of motions present in the scene are determined, and then the corresponding objects are identified based on the motion information [24], [39]. Another merging strategy is to determine the ensemble of motions and the objects in the scene simultaneously [1].

Bottom-up approaches have several advantages over top-down approaches. First, the extraction of a given object is not influenced by previously extracted objects. Second, the bottom-up approach ensures a complete decomposition of the scene. This is essential in applications such as video coding, where the segmentation is used as an alternative to the pixel-based image representation. However, bottom-up approaches are computationally expensive. Also, the resulting segmentation depends critically on the region similarity measure and the region-merging strategy.

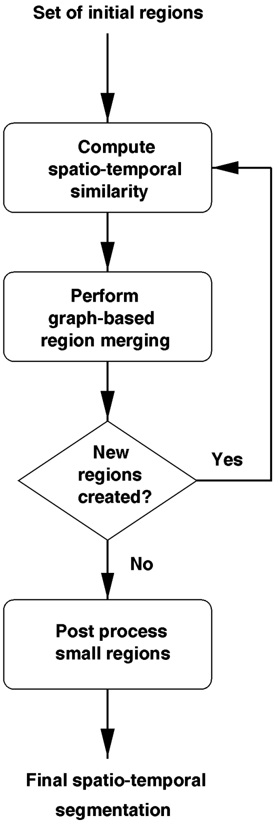

Set of initial regions
Compute spatio-temporal similarity
Perform graph-based region merging
New regions created?
Yes
No
Post process small regions
Final spatio-temporal segmentation

Fig. 1. Overview of the proposed algorithm for spatiotemporal segmentation. The set of initial regions is assumed to be known. These regions are iteratively merged together, to identify the objects present in the scene. This procedure has two phases: the computation of spatiotemporal similarities and the graph-based decision of which regions to merge. After the merging procedure terminates, a postprocessing step processes regions that are too small to represent valid objects.

3 Overview of the Proposed Technique for Spatiotemporal Segmentation

The spatiotemporal segmentation technique proposed here adopts a region-merging approach. We justify this approach by observing that most simple segmentation techniques provide an oversegmentation of the scene. Therefore, a higher-level region-merging procedure is required to identify the spatiotemporally coherent objects. Fig. 1 shows the flowchart of the proposed method. This method takes as input the set $\mathcal{R}$ of N_r initial regions that correspond, in general, to an oversegmentation of the current frame. Starting from $\mathcal{R}$, the objects constituting the scene are detected automatically. The proposed spatiotemporal segmentation is carried out in three steps. First, the spatiotemporal similarities existing between pairs of regions are computed. These spatiotemporal similarities are then used to build a graph which assimilates all the information available about the regions and their relationships. The graph is used to merge regions iteratively. At every iteration, the graph is updated to reflect the spatiotemporal characteristics of the newly formed regions. The process is iterated until no further merging occurs. In the final step, regions that are too small to represent valid objects are merged with larger objects on the basis of temporal similarity.

The proposed spatiotemporal similarity uses both spatial and temporal information efficiently. When assessing the possibility of merging two regions, the resemblances between their motion characteristics and their spatial characteristics are analyzed. The spatiotemporal similarity between the regions is expressed as a hypothesis test, the assumption to be tested being that two regions are spatiotemporally similar. The proposed measure of spatiotemporal similarity is defined as the significance level of this test. It is obtained by examining two hypotheses, one relying on spatial information and the other relying on temporal information. The former hypothesis states that two regions are spatially coherent. The significance level of this hypothesis gives their spatial similarity. The likelihood ratio test is used as the test statistic. The hypothesis test is rendered less sensitive to noise and false alarms through the use of robust data. The hypothesis based on temporal information says that the two regions have similar motion. The significance level of this hypothesis defines their temporal similarity. The hypothesis is tested through a nonparametric test statistic referred to as the *Modified Kolmogorov-Smirnov* (MKS) test. It combines the motion information present in the motion parameters as well as in the residual distributions. The test also takes into account the presence of outliers. The spatiotemporal similarity is defined as a combination of both the temporal and spatial similarities. However, more emphasis is put on the temporal similarity so as to comply with the underlying definition of an object.

The proposed region-merging strategy aims to exploit fully the information provided by the spatiotemporal similarity. The region-merging procedure is formulated as a graph-based clustering problem. The proposed strategy relies on two clustering rules, referred to as the *strong rule* and the *weak rule*. These complementary rules are designed to exploit the natural structures present in the graph, while avoiding the shortcomings and likely errors such as erroneous motion information or badly defined regions.

4 Generation of the Set of Initial Regions

We emphasize that the technique for spatiotemporal segmentation proposed here does not depend on the way in which the set of initial regions, $\mathcal{R}$, is generated. In this work, the set of initial regions is generated based on spatial, temporal, and change information. This results in spatiotemporally homogeneous regions [40].

The algorithm consists of several steps. The first step is to generate a static segmentation of the current frame which is based on the color information. Next, these regions are refined using temporal information. More precisely, their temporal homogeneity is verified. To that end, motion parameters are estimated for each region. In order to estimate the motion of the regions accurately, the effects of camera motion are first removed. Local motion estimation then takes place between the current frame and the globally motion-compensated previous frame. (The technique for motion estimation is discussed later.) The regions that have high Displaced Frame Difference (DFD) energy after compensation are further split based on color information. Note that the temporal refinement of the regions takes into account the phenomenon of disocclusion. This permits a robust evaluation of the temporal homogeneity of a given region. Finally, a change detection procedure is carried out in order to detect regions that have low contrast. The change information results from the comparison of the current frame with its prediction, which is the motion-compensated version of the previous frame.

The result of the method is the set $\mathcal{R}$ of initial regions. The regions are characterized by their spatiotemporal homogeneity. They represent all the visually important details of the image.

5 Definition of the Spatiotemporal Similarity Measure

The decision to merge two or more regions is based on an estimate of their mutual similarity. We propose a region similarity measure that exploits both spatial and temporal information. However, the measure relies more on temporal information, since we define objects primarily as coherently moving entities. For two regions, A and B, the measure of spatiotemporal similarity, $Sim(A, B)$, integrates both types of information into a single value.

5.1 The Spatial Similarity

When defining the objects forming the scene, we rely on the Gestalt "law of common fate." Therefore, the definition of the spatiotemporal similarity should rely mostly on the temporal information. However, one typically expects the regions forming an object to share some common spatial characteristics as well. When deciding whether two regions belong to the same object, the spatial information can form an important complement to the temporal information.

Spatial information has commonly been used through the property of *adjacency* [2], [25], [41], [42]. It states that two regions may be merged only if they are neighbors. This constraint intuitively makes sense, as most objects are spatially connected. Also, the adjacency requirement reduces the number of combinations of regions to be considered. However, the available spatial information is only poorly exploited by the adjacency constraint. In addition to the adjacency constraint, Thompson [37] also makes use of luminance information. The merger of two regions depends on the luminance contrast along their common border.

In this section, a measure of the *spatial similarity* between two regions is proposed. Let S_{AB} denote the spatial similarity between two regions, A and B. We consider S_{AB} to be the likelihood that the regions A and B belong to the same object as far as the spatial information is concerned. The spatial similarity ranges from 100 percent (absolute spatial similarity) to 0 percent (absolute spatial dissimilarity). S_{AB}

imposes the adjacency constraint and is set to 0 percent for regions that are not adjacent.

Consider two adjacent regions A and B. Assume the null hypothesis, H_0, that, according to the spatial information, regions A and B belong to the same object. The alternative hypothesis, H_1, states that regions A and B do not belong to the same object. As the validity of H_0 is directly proportional to its significance level, the latter is taken to be the spatial similarity S_{AB}. Formally, the hypothesis test is written as

$$\begin{cases} H_0: & \text{regions A and B are spatially similar} \\ H_1: & \text{regions A and B are spatially not similar} \end{cases} \qquad (1)$$

Many types of spatial data, such as shape information, color, and texture information, can be used to test the hypothesis described by (1). In our case, we base the test on spatial data that are representative of the luminance contrast. The premise is that if a strong contrast exists between A and B, the regions are less likely to belong to the same object. Moreover, the information should be robust in the presence of noise. In order to fulfill these requirements, the spatial information present on the common border between regions A and B is used to verify the above-mentioned hypothesis.

We consider the medians, l_{AB} and l_{BA}, of the luminance values of the two regions along their common border. The measurements l_{AB} and l_{BA} are seen as trials of two random variables, L_{AB} and L_{BA}, respectively. Both random variables are assumed to be modeled by the same Gaussian distribution, with mean, μ, and standard deviation, σ. Formally, $L_{AB} \sim N(\mu, \sigma)$ and $L_{BA} \sim N(\mu, \sigma)$. The measure of spatial similarity turns out to be the difference between the two medians, normalized by σ. In practice, the values of μ and σ are estimated over the ensemble of luminance medians for all the combinations of adjacent regions in the scene. This allows us to introduce the spatial activity existing in the entire image into the evaluation of S_{AB}.

For this hypothesis test, we choose the likelihood ratio test as the test statistic, Q_s. This test is simple and robust. It is defined as the difference between L_{AB} and L_{BA}, $Q_s = L_{AB} - L_{BA}$. This implies that Q_s is a Gaussian random variable with zero mean and a standard deviation equal to $\sqrt{2}\sigma$ (i.e., $Q_s \sim N\left(0, \sqrt{2}\sigma\right)$). The hypothesis test defined by (1) is rewritten as

$$\begin{cases} H_0: & Q_s = 0 \\ H_1: & Q_s \neq 0 \end{cases} \qquad (2)$$

The hypothesis testing thus reduces to checking whether the mean of Q_s is zero. The realization, q_s, of the test statistic, Q_s, is given by $q_s = l_{AB} - l_{BA}$. Defined as the significance level of the hypothesis test, the spatial similarity S_{AB} may now be written as

$$S_{AB} = 1.0 - \frac{1}{\sqrt{2\pi\sigma}} \int_{-q_s}^{q_s} e^{-\frac{1}{\sqrt{2\sigma^2}} x^2} dx = 1.0 - \frac{1}{\sqrt{2\pi}} \int_{-\frac{q_s}{\sigma}}^{\frac{q_s}{\sigma}} e^{\frac{-t^2}{2}} dt, \quad (3)$$

Fig. 2 shows S_{AB} as a function of the normalized value of the test statistic, $\frac{q_s}{\sigma}$. The spatial similarity decreases sharply as the value of the test statistic increases in magnitude and becomes negligible when the test statistic has values larger than three times the standard deviation, σ. This behavior is characteristic of a Gaussian distribution which is taken as the underlying statistical model of the medians of the luminance values.

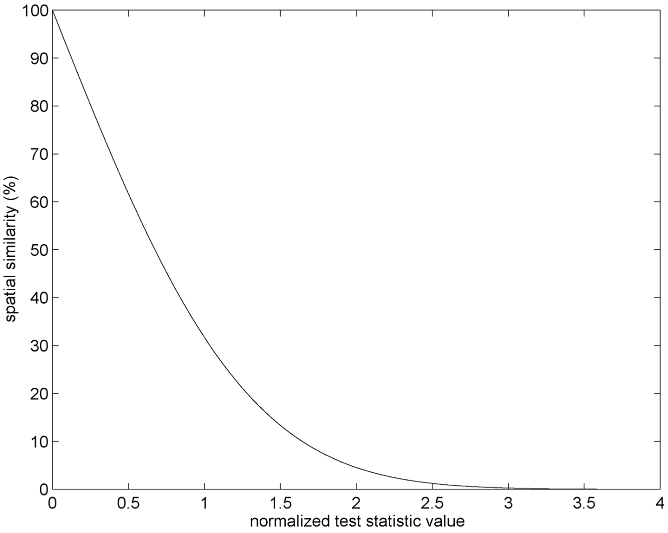


Fig. 2. The spatial similarity as a function of $\frac{q_s}{\sigma}$.

The spatial similarity S_{AB} is very robust, primarily due to the use of the median estimator. The value of S_{AB} takes into account the spatial activity existing in the entire scene. This is due to the fact that the standard deviation of Q_s incorporates information from the whole image. Moreover, the spatial similarity is symmetric ($S_{AB} = S_{BA}$) and reflexive (S_{AA} = 100 percent).

5.2 The Temporal Similarity

Temporal information plays a fundamental role in the decomposition of dynamic scenes. Indeed, under the assumption of rigid motion, a moving object is defined primarily by its temporal coherence. Thus, temporal information forms a reliable basis for deciding whether two regions belong to the same object. The similarity between the motion characteristics of two regions is referred to as the *temporal similarity* between them.

In this section, a measure of the temporal similarity, T_{AB}, between two regions, A and B, is proposed. T_{AB} ranges from 100 percent (perfect temporal similarity) to 0 percent (absolute temporal dissimilarity). To keep the problem computationally tractable, the temporal similarity is estimated only for adjacent regions. It is set to 0 percent for pairs of regions that are not adjacent.

Again, hypothesis testing is used to determine the temporal similarity between two regions. Assume that regions A and B are adjacent. The null hypothesis, H_0, is that region A moves in the same way as region B. The alternative hypothesis, H_1, states that region A does not undergo the same motion as region B. T_{AB} is defined as the significance level of H_0. Formally, the hypothesis test is written as

$$\begin{cases} H_0: & \text{the region A is temporally similar to region B} \\ H_1: & \text{the region A is temporally not similar to region B} \end{cases}$$

To test the hypothesis described in (4), we propose a modification of the well known Kolmogorov-Smirnov (K-S) test [43], called the MKS test [40]. The associated test statistic, Q_{MKS}, combines the motion information contained in the parametric representation with the motion information present in the residual distribution. The residual distribution represents the discrepancy between the motion information characterized by the motion parameters and the true optical flow of the region. Consequently, the information represented by the residual distribution is complementary to the information present in the motion parameters. Q_{MKS} takes into account the likely presence of outliers. It has been developed specifically for estimating the temporal similarity between regions.

The test statistic Q_{MKS} relies more on the residual distributions than on the parametric characterization of the motion information. A given optical flow may be characterized by a set of motion parameters that does not represent the motion accurately. This is especially true for regions where the motion is not very well-defined. Thus, the motion parameters alone cannot be considered very reliable. On the other hand, two residual distributions, generated using different sets of motion parameters for the same region, will both reflect almost the same motion characteristics. Therefore, the residual distributions are assumed to provide very reliable information.

Consider two regions, A and B, with motion parameters $\vec{M}_A$ and $\vec{M}_B$, respectively. For region A, we compute two residual distributions, h_1 and h_2. The distribution h_1 is obtained by compensating A for motion using the parameters $\vec{M}_A$, while h_2 is obtained when A is motion compensated with the motion parameters $\vec{M}_B$. Based on these definitions, the realization, q_{MKS}, of the test statistic Q_{MKS} is expressed as

$$q_{MKS} = \lambda \max_x \left| F_1(x) - F_2(x) \right|, \qquad 0 \le \lambda \le 1, \tag{5}$$

where

$$\begin{cases} F_1(x) & = \int_{-\infty}^{x} w(x) h_1(x) dx \\ F_2(x) & = \int_{-\infty}^{x} w(x) h_2(x) dx \end{cases} \tag{6}$$

are the weighted cumulative residual distributions corresponding to the residual distributions h_1 and h_2, respectively. The function $w(x)$ is called the *weighting function* and is a positive $\mathcal{L}_2$ function such that $w(x) \in [0, 1]$, $\forall x$. λ is called the *pondering factor*.

The test statistic Q_{MKS} is nonparametric and uses the residual distribution as a whole to perform the hypothesis testing. The underlying idea is to measure the maximum discrepancy existing between the two cumulative distributions $F_1(x)$ and $F_2(x)$. The test simply tries to decide whether $F_1(x)$ and $F_2(x)$ correspond to the same (unknown) random process. This is in sharp contrast to techniques which reduce the distribution information to a single parameter. Not only does the use of a single parameter limit the amount of information being exploited in the test, but it also runs the risk of poor results due to incorrect estimation of parameters. The nonparametric nature of Q_{MKS} renders the test very robust and flexible, since no predefined model need be assumed for the distributions.

Compared to the K-S test, the proposed test statistic, Q_{MKS}, differs in two ways. The first modification is the introduction of the weighting function, $w(x)$. This function is used to reduce the influence of outliers on the test. We will discuss it in detail in Section 5.2.1. The second modification is the *pondering factor* λ, which supplies to the test the motion information embodied in the motion parameters. This factor directly affects the discrepancy found between the cumulative distributions $F_1(x)$ and $F_2(x)$. It is discussed in more detail in Section 5.2.2.

Using the test statistic, Q_{MKS}, we can formalize the hypothesis test for the temporal similarity (4) as

$$\begin{cases} H_0: & Q_{MKS} = 0 \\ H_1: & Q_{MKS} > 0 \end{cases} \tag{7}$$

The temporal similarity T_{AB} is a function of both the test statistic value, q_{MKS}, and the area of region A, $Area(A)$, and is given by [44]:

$$T_{AB} = 2 \sum_{j=1}^{\infty} \left((-1)^{j-1} e^{-2j^2 \delta^2} \right), \tag{8}$$

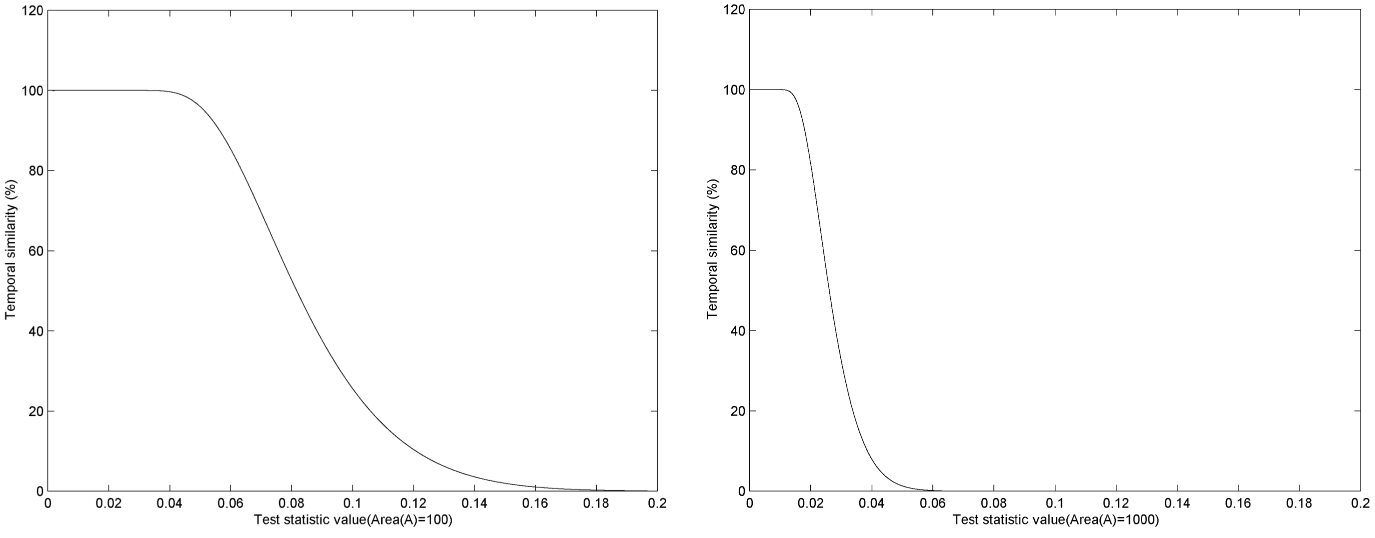
where

$$\delta = \left(\sqrt{Area(A)} + 0.12 + \frac{0.11}{\sqrt{Area(A)}} \right) q_{MKS}.$$

Fig. 3 shows examples of T_{AB} as a function of q_{MKS} for different values of $Area(A)$. The typical curve features a sharp transition which approaches the origin as the value of $Area(A)$ increases. For a given value of q_{MKS}, the larger the value of $Area(A)$, the lower the temporal similarity. This behavior makes intuitive sense, since if two distributions are the same, the discrepancy between them should diminish as the number of trials (i.e., $Area(A)$ in this case) increases. Note that the temporal similarity as defined by (8) is not symmetric, that is, $T_{AB} \neq T_{BA}$. This property of directionality proves fundamental in the region-merging process. The similarity is reflexive, that is, T_{AA} = 100 percent.

5.2.1 The Weighting Function w(x)

The weighting function, $w(x)$ (in (6)), has been introduced to make the test statistic Q_{MKS} robust in the presence of outliers. Its purpose is to modulate the importance of discrepancies between the distributions depending on the value, x, of the residue. In particular, the weighting function $w(x)$ renders the hypothesis-testing procedure coherent with a robust motion estimation procedure. By definition, robust motion estimators are less sensitive to data outliers. They do not attribute the same importance to the different residues [18]. The residues close to zero are typically considered more significant than larger residues in order to decrease the influence of outliers in the estimation process. Consider the case where an M-estimator [45] is used. Given an M-estimator $\rho(x)$, the relative importance of the residue, x, is determined by the value of the weight function, $\frac{\dot{\rho}(x)}{x}$



(a)

(b)

Fig. 3. The temporal similarity as a function of the q_{MKS} test statistic value with, respectively, (a) *Area*(*A*) = 100 and (b) *Area*(*A*) = 1,000. The two plots have the same scale.

[18]. The weighting function, $w(x)$, is the normalized version, in the range [0,1], of $\frac{\dot{\rho}(x)}{x}$. For instance, consider the case when motion estimation is performed with the Geman-McClure estimator [33], $\rho_{GM}(x)$, where

$$\rho_{GM}(x) = \frac{\frac{x^2}{\sigma^2}}{1 + \frac{x^2}{\sigma^2}}, \tag{9}$$

with σ denoting the scale factor. The corresponding weighting function, $w(x)$, is

$$w(x) = \frac{1}{\left(1 + \frac{x^2}{\sigma^2}\right)^2}. \tag{10}$$

Clearly, the function $w(x)$ favors low residues, reaching the maximum weight of unity for the null residue ($x = 0$). As the residue becomes larger, its influence gradually decreases until it becomes insignificant, with a weight close to zero.

5.2.2 The Pondering Factor λ

The pondering factor λ is used to introduce the motion information existing in the motion parametric representation, into the Q_{MKS} test statistic. Its role is to ponder the discrepancy found between the two weighted cumulative distributions $F_1(x)$ and $F_2(x)$.

The definition of λ relies on the distance between the n motion parameters of the region A, $\vec{M}_A$, and those of region B, $\vec{M}_B$. More precisely, λ is derived from the significance level, $\alpha_p \in [0, 1]$, of the hypothesis that the parameter vectors $\vec{M}_A$ and $\vec{M}_B$ are the same. For a given α_p,

$$\lambda = 1 - \alpha_p. \tag{11}$$

Let us examine the two limiting situations.

- In case $\vec{M}_A$ is similar to $\vec{M}_B$, $\alpha_p \approx 1$, implying that $\lambda \approx 0$. According to (5), q_{MKS} decreases significantly, irrespective of the discrepancies present between the residual distributions. Thus, the temporal similarity T_{AB} is high.
- If $\vec{M}_A$ and $\vec{M}_B$ are very dissimilar, $\alpha_p \approx 0$. Since $\lambda \approx 1$, it has almost no influence on q_{MKS}. In this case, the determination of temporal similarity T_{AB} relies completely on the information derived from the residual distributions.

To derive the value of α_p, we again use a hypothesis test. Each motion parameter, $\vec{M}(i)$, $i = \{1, \cdots, n\}$, is modeled as a Gaussian random variable with mean μ_i and standard deviation σ_i. The corresponding test statistic, Q_i, is chosen to be the likelihood ratio test (i.e., $Q_i = \vec{M}_A(i) - \vec{M}_B(i)$. The test statistic Q_i is a normal random variable with $Q_i \sim N\left(0, \sqrt{2}\sigma_i\right)$. Consequently, the significance level, α_i, is

$$\alpha_i = 1.0 - \frac{1}{\sqrt{2\pi}\sigma_i} \int_{-q_i}^{q_i} e^{-\frac{1}{2\sigma_i^2}x^2} dx, \tag{12}$$

where q_i is set to $\vec{M}_A(i) - \vec{M}_B(i)$. In practice, the standard deviation, σ_i, is estimated over the population composed of the parameters, $\vec{M}(i)$, of all the regions in the frame.

The significance level, α_p, is defined as follows:

$$\alpha_p = \min_i \left\{\alpha_i, i = 1, \cdots, n\right\}. \tag{13}$$

This gives a very conservative estimate of the similarity between $\vec{M}_A$ and $\vec{M}_B$.

As mentioned before, the pondering factor, λ, introduces motion information present in the motion parametric representation into the Q_{MKS} test statistic. However, the definition of λ is very conservative. Therefore, the parametric information is only utilized when it is indisputably found to be reliable.

5.3 The Spatiotemporal Similarity

The moving objects forming the scene are characterized by their temporal coherence and, to a lesser extent, by their spatial homogeneity. Therefore, we require a measure of *spatiotemporal similarity* between regions.

The best way of combining the two types of information remains an open question. In this work, the spatiotemporal similarity, $Sim(A, B)$, of two regions A and B is specified as a combination of the spatial similarity, S_{AB} (see Section 5.1), and the temporal similarity, T_{AB} (see Section 5.2). The spatiotemporal similarity $Sim(A, B)$ is understood as the likelihood of region A belonging to the same object as region B. The likelihood depends on both temporal and spatial information. Recall, however, that an object is likely to be composed of regions having different spatial characteristics. Thus, $Sim(A, B)$ must rely mainly on the temporal information (i.e., T_{AB}). For any region X, let Υ_X be the ensemble of its adjacent regions. The proposed spatiotemporal similarity $Sim(A, B)$ is written as follows:

$$Sim(A, B) = T_{AB} - f_L\, T_{AB}(Max - S_{AB}), \tag{14}$$

with $0 \le f_L \le 1$, and

$$Max = \max(S_{AE}, S_{EF}),$$

$$S_{AE} = \max_{I \in \Upsilon_A}\left(S_{AI}\right),$$

$$S_{EF} = \max_{I \in \Upsilon_E}\left(S_{EI}\right).$$

Equation (14) reflects the fact that T_{AB} is the most significant term in the spatiotemporal similarity $Sim(A, B)$. S_{AB} is used just as a corrective factor. The influence of the spatial similarity is regulated by the user-defined factor f_L, referred to as the *luminance factor*. Consider the limiting cases. If the luminance factor f_L is set to zero, the spatial information S_{AB} is not used, and the spatiotemporal similarity $Sim(A, B)$ is equal to the temporal similarity T_{AB}. Conversely, the correction induced by the spatial information S_{AB} attains its maximum value when the luminance factor f_L is set to unity. At most, the spatial information may completely wipe out the temporal information, resulting in $Sim(A, B)$ being set to zero.

The magnitude of the maximum correction allowed through the term S_{AB} is a function of the term Max. This term conveys information about the general spatial coherence of region A with its neighboring regions. It is closely linked to the spatial activity existing around region A. Fig. 4 illustrates the derivation of Max. First, we determine the region E adjacent to A, $E \in \Upsilon_A$, which is spatially the most similar to region A. Next, we determine region F adjacent to E, $F \in \Upsilon_E$, which is spatially the most similar to region E. The term Max is defined as the maximum between the spatial similarities S_{AE} and S_{EF}. It represents the maximum correction the spatial similarity may make to T_{AB}. In case Max is small (i.e., close to zero), region A is by definition lying in an area of high contrast. This implies that the spatial information is not very useful and it should not be allowed to influence $Sim(A, B)$ strongly. In the case where Max is large (i.e., close to unity), the neighborhood of the region A is spatially quite homogeneous. Consequently, the spatial information S_{AB} should be allowed to play a stronger role in the definition of $Sim(A, B)$.

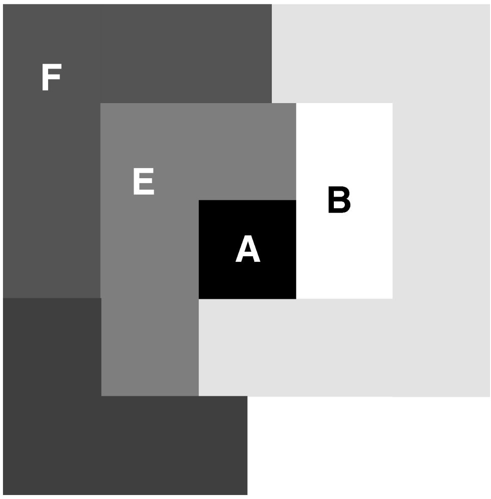


Fig. 4. Derivation of *Max*. The spatial coherence of region *A* with its neighborhood is checked. The term *Max* reflects the importance of the spatial coherence in the vicinity of region *A*.

By construction, $Sim(A, B)$ estimates the similarity between region A and region B robustly. This robustness derives from the definition of both T_{AB} and S_{AB}, as well as the definition of $Sim(A, B)$ itself. The spatiotemporal similarity is asymmetric (i.e., $Sim(A, B) \neq Sim(B, A)$) and reflexive ($Sim(A, A) = 100$ percent).

6 The Region-Merging Strategy

Spatiotemporal region merging can be seen as a clustering problem: Regions that have similar spatiotemporal characteristics are to be clustered together. After defining the spatiotemporal similarity measure, we need a strategy to use this information in order to form clusters of regions. In the relevant literature, two types of strategies have been proposed. Approaches of the first type start by identifying the different motions present in the scene. Then, the regions that correspond to each motion are merged to form an object [3], [24], [39]. Other approaches perform the motion estimation and region merging simultaneously. Each region is tentatively merged with each of its neighboring regions, in turn. A particular merger is accepted in case the region similarity satisfies a predefined criterion [1], [2], [25]. The strategy used in this work falls in the second category. The merging of the regions relies on a graph representation. Two complementary graph-clustering rules are presented: the strong rule and the weak rule. These rules are applied successively and lead to a robust definition of the objects in the scene.

6.1 Graph-Based Region Clustering

The region-merging process may be concisely formulated as a graph-based clustering problem. A graph, G, is used to represent the information upon which the region-merging

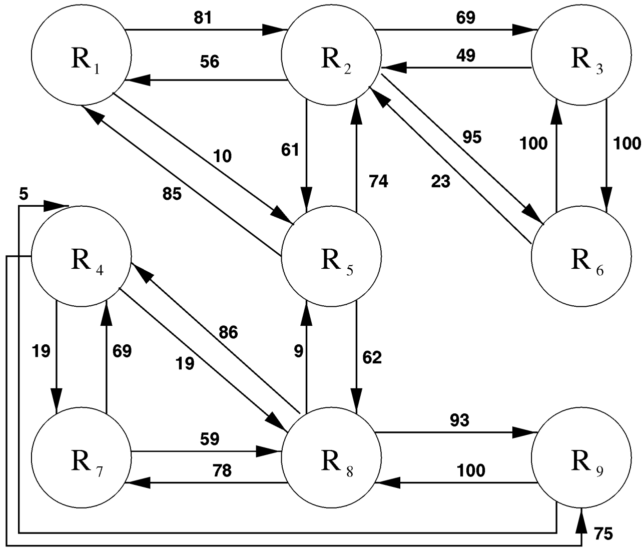


Fig. 5. Graph representation of the region similarities. The vertices represent regions, and the edges represent the spatiotemporal similarities between the regions. The similarities are expressed as percentages.

process is based. Each vertex in G represents one of the N_r regions in the set $\mathcal{R} = \left\{R_1, R_2, \cdots, R_{N_r}\right\}$. Every edge represents the spatiotemporal similarity between the two regions corresponding to the two vertices it connects. The set of edges is denoted $\mathcal{A}$. The merging strategy should exploit the structures revealed by the graph representation.

Several strategies have been proposed for graph clustering. A good introduction to the topic is given in [46]. Most graph-based clustering algorithms are iterative in nature. They differ mainly in the way the vertices to be merged in a given iteration are selected. However, most of the theory for graph-based clustering has been developed under the assumption of a symmetrical similarity measure. That is, the similarity of vertex A to vertex B is assumed to be the same as the similarity of B to A. This implies that the graph has undirected edges, and each pair of vertices, if connected, are attached by a single edge. Thus, in the context of spatiotemporal region merging, most graph-based techniques rely on symmetric similarity measures [25], [42], [47]. The usual strategy follows the Greedy Merging Algorithm (GMA). This algorithm iteratively merges the two regions showing the strongest similarity, until a stopping criterion is triggered.

The similarity measure defined in Section 5.3 is not symmetric: $Sim(A, B) \neq Sim(B, A)$. Clearly, therefore, the corresponding graph is weighted as well as directed. Pairs of vertices in the graph are connected by two directed edges. A hypothetical example of the graph G is depicted in Fig. 5. The set of regions $\mathcal{R}$ includes nine regions, $\{R_1, \cdots, R_9\}$. The spatiotemporal similarity of region A with region B is represented with an arrow going from the vertex B to vertex A. For instance, the spatiotemporal similarity of region R_2 with region R_1 is 81 percent. Recall that the similarity is set to 0 percent for regions that are not adjacent. These null edges as well as the edges related to the reflexive property are not represented in the graph.

Graph G contains all the information necessary for the region-merging process. The merging process corresponds to the extraction of clusters existing in the graph. In order to accomplish this, graph-based clustering rules have to be defined. Their derivation is closely linked to the graph at hand and to the problems which may have occurred while constructing it. The directional and weighted edges define natural structures in G which are directly related to the objects forming the scene. In the example shown in Fig. 5, three natural structures may be easily identified. They are, respectively, $\{R_1, R_2, R_5\}$, $\{R_3, R_6\}$, and $\{R_4, R_7, R_8, R_9\}$. The graph-clustering rules aim to exploit this structural information. The rules also take into account the inaccuracies and errors reported in the graph. In our case, certain problems stem from regions in $\mathcal{R}$ that have badly estimated motion information. These are typically small regions, created due to noise. The incorrect motion information may jeopardize the clustering process by corrupting the set $\mathcal{A}$ of edges. Another type of problem arises from erroneous motion parameters. Spurious edges are produced, which may also lead to undesirable clustering.

We propose two clustering rules that take the above considerations into account. They are called the *strong rule* and the *weak rule*, respectively. They are mutually complementary and are designed to address the ensemble of different situations which may occur during the merging process. In order to decrease the complexity of the task, both rules operate on a thresholded graph, which is obtained by deciding whether to accept or reject the hypothesis of spatiotemporal similarity between pairs of regions. Given a threshold of acceptance, t, the hypothesis that region A is spatiotemporally similar to region B is accepted if and only if

$$Sim(A, B) \geq t. \tag{15}$$

In case of acceptance (and using the arrow notation as in Fig. 5), we write $(B \rightarrow A)$. The notation $(B \rightarrow A)$ represents a binary-valued relationship, implying that the spatiotemporal characteristics of B can also be used to describe A. (The absence of an arrow represents the opposite case.) Fig. 6 gives an example of a thresholded graph. This has been generated by setting $t = 60$ percent on the weighted graph given in Fig. 5.

Next, the two clustering rules are explained. The strong rule identifies the natural structures present in the graph. The weak rule assumes that among the regions of $\mathcal{R}$, some are very well-defined, both spatially and temporally. The weak rule uses them as seeds to determine which regions should be merged. In the following, N_c denotes the number of clusters and is initially set to N_r. Through clustering, the regions are distributed into clusters C_i, $i \in [1, \cdots, N_c]$, $N_c \leq N_r$. The number of clusters N_c is automatically determined by the clustering rules. Below, we elaborate on the two rules. The way these rules are used is discussed later.

6.2 The Strong Rule

Graph G is constructed from sets $\mathcal{R}$ and $\mathcal{A}$. This implies that the presence of inaccuracies or errors in these sets may

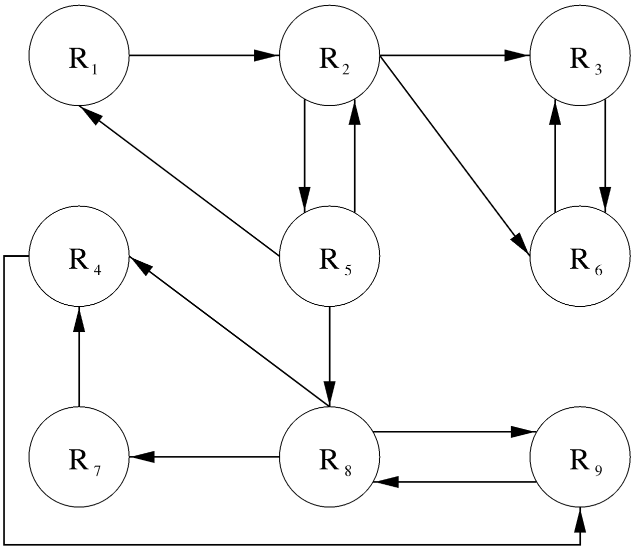


Fig. 6. Example of a thresholded graph. The relationships between the vertices are binary valued. Thus, ($R_i \to R_j$) implies that the spatiotemporal properties of R_i also suit R_j very well.

severely influence the graph-based region-merging process. The graph-based clustering procedure should thus concentrate only on the secure pieces of information. Only regions with well-defined and correctly estimated motion should be considered. In view of these considerations, the strong rule is designed to cluster regions which form a cycle in the graph. More precisely, the strong rule clusters regions belonging to the same cycle only.

The strong rule operates on a binary graph. Such a graph is generated by thresholding a weighted graph using a predefined threshold, t_{sr}. The cycle requirement imposed by the strong rule defines the clusters C_i, $i \in [1, \cdots, N_c]$, as follows:

$$C_i = \{R_j \mid \exists (R_k, R_l \in C_i, k \neq j, l \neq j)$$

such that

$$(R_j \to R_l \text{ and } R_k \to R_j), \qquad (16)$$

with R_j, R_k, $R_l \in \mathcal{R}$.

By looking for cycles in the graph, the strong rule focuses on the secure information, thus avoiding the problems arising from doubtful regions or erroneous motion parameters. The cycle implies that each region in the cluster is spatiotemporally similar to at least one other region of the cluster. In this way, only regions with strong spatiotemporally similarity are clustered together. The cycle requirement avoids regions with wrongly estimated motion parameters. Also, the cycle requirement excludes the doubtful regions for which the motion characteristics are not well-estimated. If a region is not assigned to any of the clusters defined by (16), it is regarded as a separate cluster.

The strong rule imposes severe constraints on the clustering process. Thus, the clusters are only formed of regions which are very likely to belong to the same object. Two types of clusters result from the application of the strong rule. One type consists of clusters which contain several very well-defined regions. These clusters foreshadow the objects forming the scene. The remaining clusters typically contain isolated (initial) regions which are either oddly defined or for which the motion information is spurious. These regions are processed using the weak rule.

6.3 The Weak Rule

The weak rule is very well-suited to cases where badly defined or small regions coexist with regions having a strong semantic significance. The weak rule merges the former regions with the most appropriate of the latter ones. The weak rule generalizes the GMA [42] to cluster regions hierarchically. This is achieved by iteratively thresholding the graph with successively lower thresholds until the lowest predefined threshold t_{wr} is reached. At each iteration, the current clustering relies on the clusters found in the previous iteration.

The threshold starts at 100 percent and decreases, by a fixed step, to its lowest allowed value, t_{wr}. In each iteration, the clusters previously defined serve as the basis for the current clustering process. This implies that the number of clusters, N_c, decreases as the hierarchical clustering progresses. The hierarchical approach allows *nondynamic merging* of regions. The merging procedure deals with clusters of regions. This utilizes the structural information in the graph. This is in contrast with the GMA which deals with only individual regions. Although clusters of regions are created, there is no need to merge them and update the graph. This improves the robustness of the technique and also reduces the computational complexity.

In each iteration, the weak rule determines which regions should be merged. This is performed in two steps. For each cluster, C_m, $m \in [1, \cdots, N_c]$, the weak rule first determines the set Ω of clusters C_i, $i \in [1, \cdots, N_c]$ and $i \neq m$ with which C_m could be merged. For each cluster, C_i, the number of edges from C_i to C_m is determined. This number is compared to the number of regions present in C_m. The set Ω is defined as

$$\Omega = \left\{ C_i, i \neq m \,\middle|\, \sum_{R_k \in C_i} \sum_{R_l \in C_m} \left(R_k \to R_l\right) \geq \mathrm{Card}\left(C_m\right) \right\} \qquad (17)$$

where Card(C_m) denotes the cardinality of the cluster C_m, and R_k, $R_l \in \mathcal{R}$. Three cases are possible.

1) The set Ω may be empty. In this case, the cluster C_m remains as before.
2) The set Ω may contain a single element C_i. In this case, this element is merged with C_m.
3) In case the set Ω has several elements, further tests are required in order to select the cluster $C_s \in \Omega$ with which the cluster C_m should be merged. This selection is carried out according to the following rule

$$C_s = \left\{ \max_{\phi}\left(\max_{\pi}\left(\max_{\chi}\left(C_i \in \Omega\right)\right)\right)\right\}, \qquad (18)$$

where

$$\chi = \sum_{R_k \in C_i} \sum_{R_l \in C_m} \left(R_k \to R_l\right),$$

$$\pi = Area\left(C_i\right),$$

$$\phi = \sum_{R_k \in C_i} \sum_{R_l \notin C_i} \left(R_k \to R_l\right),$$

and $Area(C_i)$ is the area of the cluster C_i.

According to (18), the process of selecting cluster C_s involves several stages. First, the number of edges, χ, from each cluster $C_i \in \Omega$ to cluster C_m is determined. The cluster, C_s, with the largest value of χ is merged with C_m. In this case, cluster C_s is indeed the cluster whose regions are spatiotemporally the most similar to the regions of cluster C_m. If two or more elements of Ω have the same value of χ, a further selection is made on the basis of the area, π, of the cluster. The cluster, C_s, with the largest area is selected. This selection is based on the observation that the likelihood of having erroneous motion estimation decreases with the area. For the same value of χ, the cluster with the largest area should therefore be chosen as the cluster C_s. Finally, two or more clusters in Ω may have the same values for χ and π. In this case, the tie is broken by examining the number of edges, ϕ, from each cluster $C_i \in \Omega$ to all other clusters. The cluster, C_s, with the largest value of ϕ is selected. This choice has a very simple reason. As a cluster gathers new regions, it tends to isolate itself from the other clusters. In other words, when the cluster representative of an object has gathered all the regions forming this object, the cluster has mostly internal edges, with only a few (probably erroneous) external edges remaining. Conversely, a cluster, which has not yet collected all the regions corresponding to the object it represents, tends to have many edges pointing to other clusters. Clusters of the latter type are favored by the selection based on the value of ϕ.

6.4 Application of the Region-Merging Rules

Now we present the region-merging strategy used in this work. An overview of the adopted region-merging strategy is shown in Fig. 7. The strong rule is applied first. Indeed, we consider the natural structures existing in the graph as prime information. Most of the regions resulting from the strong rule are reasonable first approximations of the objects present in the scene. There may be a few regions with erroneously estimated (or otherwise ill-defined) motion information. These regions are processed by the weak rule which attempts to merge them with the other well-defined regions. For the graph shown in Fig. 6, the two clustering rules act as follows. The strong rule defines the three clusters $\{R_1, R_2, R_5\}$, $\{R_3, R_6\}$, and $\{R_4, R_7, R_8, R_9\}$. After applying the weak rule, the clustering procedure results in the clusters $\{R_1, R_2, R_3, R_5, R_6\}$ and $\{R_4, R_7, R_8, R_9\}$. For both the strong and the weak rules, the merging process is carried out iteratively and involves a dynamic update of the graph. At each iteration, first the graph is thresholded, then the regions are merged, and finally the graph is updated. Initially set to 100 percent, the threshold value for the strong rule, t_{sr}, is recomputed after each iteration. To that end, the maximum value E_s that would still allow the strong rule to carry out a merging is determined from the edges in the graph. The threshold t_{sr} is then defined as

$$t_{sr} = Int(E_s) - D_{sr}, \tag{19}$$

where D_{sr} is a predefined step size and the function $Int(\cdot)$ returns the integer part of its argument. At the end of each iteration, the graph is updated by recomputing the temporal and spatial characteristics of the newly created regions and then recomputing the similarities among the current set of regions. The iterative procedure stops either when t_{sr} is less than a predefined value t_{lsr} or if no merging occurs in the current iteration.

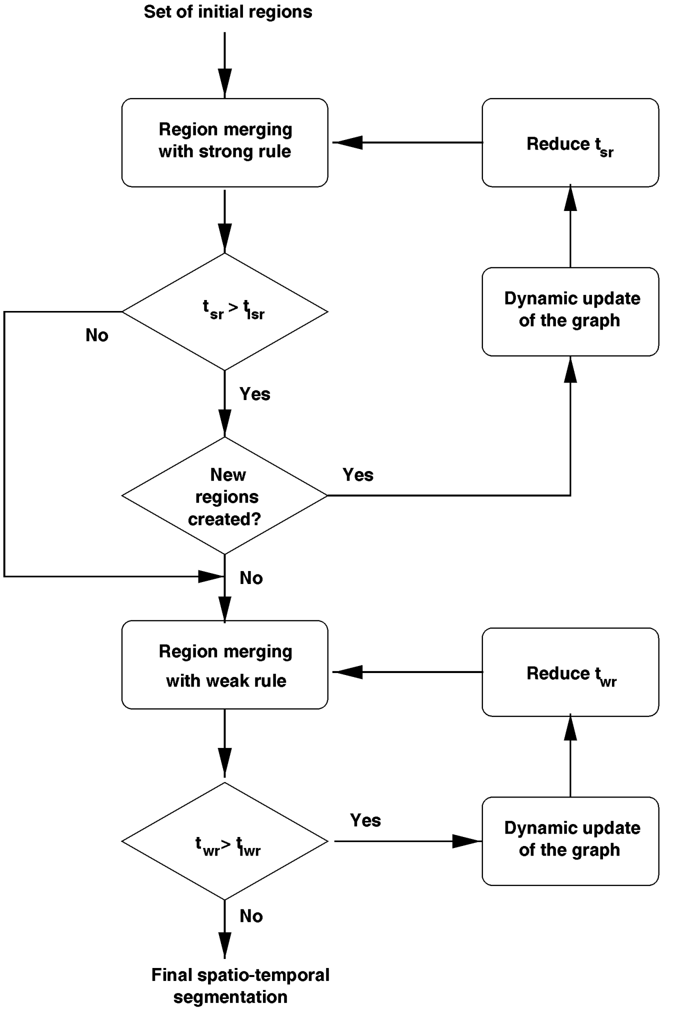

Set of initial regions
Region merging with strong rule
Reduce t_{sr}
$t_{sr} > t_{lsr}$
No
Yes
Dynamic update of the graph
New regions created?
Yes
No
Region merging with weak rule
Reduce t_{wr}
$t_{wr} > t_{lwr}$
Yes
Dynamic update of the graph
No
Final spatio-temporal segmentation

Fig. 7. The region-merging strategy. The set of initial regions is assumed to be provided. First, the strong rule is applied iteratively, and the graph is updated to reflect the regions identified in each iteration. When no further merging occurs, the weak rule is applied iteratively.

A similar strategy is used for the weak rule. In each iteration, the graph is first thresholded using a threshold t_{wr} which is computed (for every iteration) as follows:

$$t_{wr} = Int(E_w) - D_{wr}, \tag{20}$$

where D_{wr} is a predefined step size for lowering t_{wr}, and E_w is the maximum value among the edges in the graph that would still allow the weak rule to merge some regions. The merging process stops when t_{wr} is lower than a predefined value t_{lwr}.

7 Experimental Results

Some results of using the proposed technique for spatiotemporal segmentation are presented in this section. We show one set of results for three video sequences that are well known in the video-coding community, namely, “Akiyo,” at a frame rate of 10 Hz, “Table Tennis,” at 25 Hz, and “Foreman,” also at 25 Hz. As mentioned before, the proposed technique takes as input an initial set of regions which represents an oversegmentation of the scene. Then, the regions are iteratively merged in a two-stage process in order to produce the final spatiotemporal segmentation. The figures in this section show the previous and current frames and the set of initial regions used to obtain the results. The intermediate results obtained after the first stage (the strong rule) and the final spatiotemporal segmentation (after application of the weak rule) are also presented. Region boundaries from the final segmentation are also shown superimposed on the frame under consideration.

The shapes of the initial regions are strongly reflected in the final outcome. For example, if the initial regions are rectangular blocks, the final segmentation will consist of regions with blocky boundaries. Nevertheless, the proposed region-merging technique is applicable to initial regions generated using any method. We present results using sets of initial regions generated using the method described in Section 4. Results using blocky initial regions are presented in [40].

When estimating the motion of the regions, a two-stage global/local motion estimation approach is used [48]. Both global and local motions are modeled as affine motions and are estimated using the Minimum Absolute Difference (MAD) estimator. Global motion estimation relies only on the background information and is performed using a matching technique. The parameters of global motion are used to compensate for the camera motion, leading to a better definition of the local motions. The local motions are estimated between the globally affine motion compensated version of the previous frame and the current frame [3]. The local motion parameters are estimated using a matching technique. The computational complexity is reduced by building a Gaussian pyramid of the input images. This allows a nonexhaustive search while avoiding local minima. The final motion parameters at one level of the pyramid propagate as initial estimates to the next level. At each level of the pyramid, an iterative deterministic relaxation scheme is applied during the estimation procedure, to avoid local minima. The set of motion parameters of a given region is compared to the sets of motion parameters of its neighboring regions. The set of parameters providing the lowest prediction error is chosen as the new set of parameters for the region. This relaxation process is repeated until no further change occurs.

In order to perform the spatiotemporal segmentation, the weighting function $w(x)$ must be determined (see Section 5.2.1). Based on the weight function of the MAD motion estimator, we have

$$w(x) = \left|\frac{1}{x}\right|. \tag{21}$$

However, this weighting function cannot be used due to the discontinuity it presents at $x = 0$. We avoid this problem by defining $w(x)$ as follows:

$$w(x) = \frac{1}{1+|x|}. \tag{22}$$

The value of the luminance factor, f_L, in (14) provides information about the relative significance of spatial and temporal information. The values of t_{lsr}, t_{lwr}, and f_L used for the various sequences in our experiments are reported in Table 1.

TABLE 1
THE SETTINGS UTILIZED FOR THE THREE VIDEO SEQUENCES USED IN THE EXPERIMENTS REPORTED HERE

Variable	Akiyo	Table Tennis	Foreman
t_{lsr}	0.1	10.0	0.1
t_{lwr}	0.1	25.0	0.7
f_L	0.1	0.1	0.3

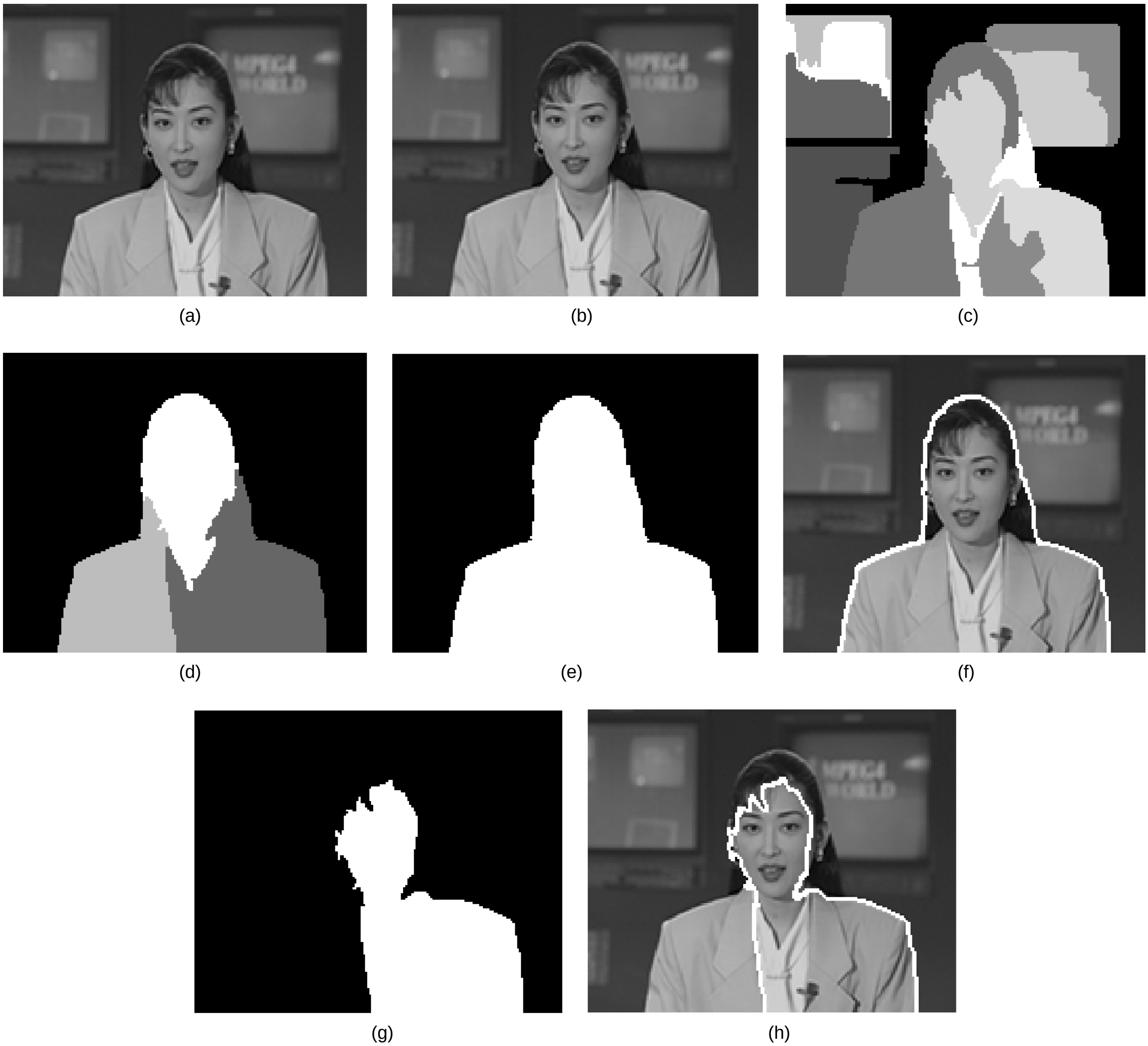
7.1 Results Using Spatiotemporally Homogeneous Initial Regions

In this section, we present spatiotemporal segmentation results for the three sequences: “Akiyo,” “Table-Tennis,” and “Foreman.” The set of initial regions is generated using the technique presented in Section 4. By construction, these regions are spatiotemporally homogeneous.

The results of the proposed technique for spatiotemporal segmentation are compared with those obtained using the technique proposed by Dufaux et al. [3]. Their technique merges the regions into objects by clustering the motion parameters of the various regions in the initial set, using the k-medoid clustering algorithm [49]. This is a supervised technique and requires the user to specify the expected number of objects beforehand.

Fig. 8 shows the experimental results obtained for a frame of the “Akiyo” sequence. In this sequence, the body of Akiyo is not absolutely rigid, and different areas have different kinds of motion (for example, the head tilts and the eyes close simultaneously). Also, the color contrast between Akiyo’s hair and the background is not very strong. The corresponding threshold values (see Table 1) are chosen so that the entire body of Akiyo is segmented out as one object.

The previous frame (after global motion compensation) and the current frame are shown in Figs. 8a and 8b, respectively. The set of initial regions is shown in Fig. 8c. It contains 14 regions. The application of the strong rule reduces this to four regions. The corresponding intermediate result is shown in Fig. 8d. The weak rule processes this intermediate segmentation to create two final regions. Fig. 8e shows the final segmentation, and Fig. 8f shows the boundaries of this segmentation superimposed onto the current frame. Note that Akiyo has been segmented perfectly, including the hair on her head, which, in some places, has spatial characteristics very similar to those of the background. Fig. 8g shows the spatiotemporal segmentation of the current frame obtained using the technique proposed by Dufaux et al. [3], and Fig. 8h shows the corresponding region boundaries superimposed onto the current frame. This segmentation has been generated by requiring the technique to segment the image into two regions. As can be seen, parts of Akiyo’s body are merged with the background. These are regions of low motion (which are nevertheless in sufficiently strong contrast to the background). The technique proposed in this paper clearly outperforms the method used by Dufaux et al.


(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 8. “Akiyo”: Spatiotemporal segmentation using arbitrarily shaped initial regions. (a) Previous frame after global motion compensation. (b) Current frame. (c) Set of initial regions (14 regions). (d) Intermediate spatiotemporal segmentation obtained from the strong rule (four regions). (e) Final spatiotemporal segmentation obtained after the weak rule (two regions). (f) Boundaries of the final segmentation superimposed onto the current frame. (g) Spatiotemporal segmentation obtained using the technique of Dufaux et al. by requiring a two-region partition. (h) Boundaries obtained by Dufaux’s method superimposed onto the current frame.

In Fig. 9, the different stages of the spatiotemporal segmentation for one frame of the “Table Tennis” sequence are shown. (The thresholds used for this experiment are given in Table 1.) In this sequence, all the objects in the scene have fairly distinct motion and luminance characteristics. Therefore, the threshold values should not be very low. Indeed, if we wanted to have the entire right arm (with racquet) as one object, the threshold values would have to be lowered. Fig. 9a shows the global motion-compensated version of the previous frame used. The current frame is shown in Fig. 9b. Fig. 9c shows the initial segmentation, containing 31 regions. The strong rule merges these regions to produce a segmentation containing 11 regions (see Fig. 9d). Fig. 9e shows the final segmentation, containing five objects, obtained after applying the weak rule to the results of the strong rule. The five objects are: the background, the arm, the ball, the hand with the racquet, and the left hand. The spatiotemporal segmentation from the method of Dufaux et al. is produced by requiring a segmentation of five objects. However, even this additional information does not produce a satisfactory spatiotemporal segmentation (see Fig. 9g and 9h). Some parts of the arm are merged with the table, while other parts are merged with the background.

The spatiotemporal segmentation results for a typical frame of the “Foreman” sequence are given in Fig. 10. The low threshold values for the two clustering rules encourage more regions to be merged. Also, the luminance information for Foreman’s face is highly distinctive. Therefore, this

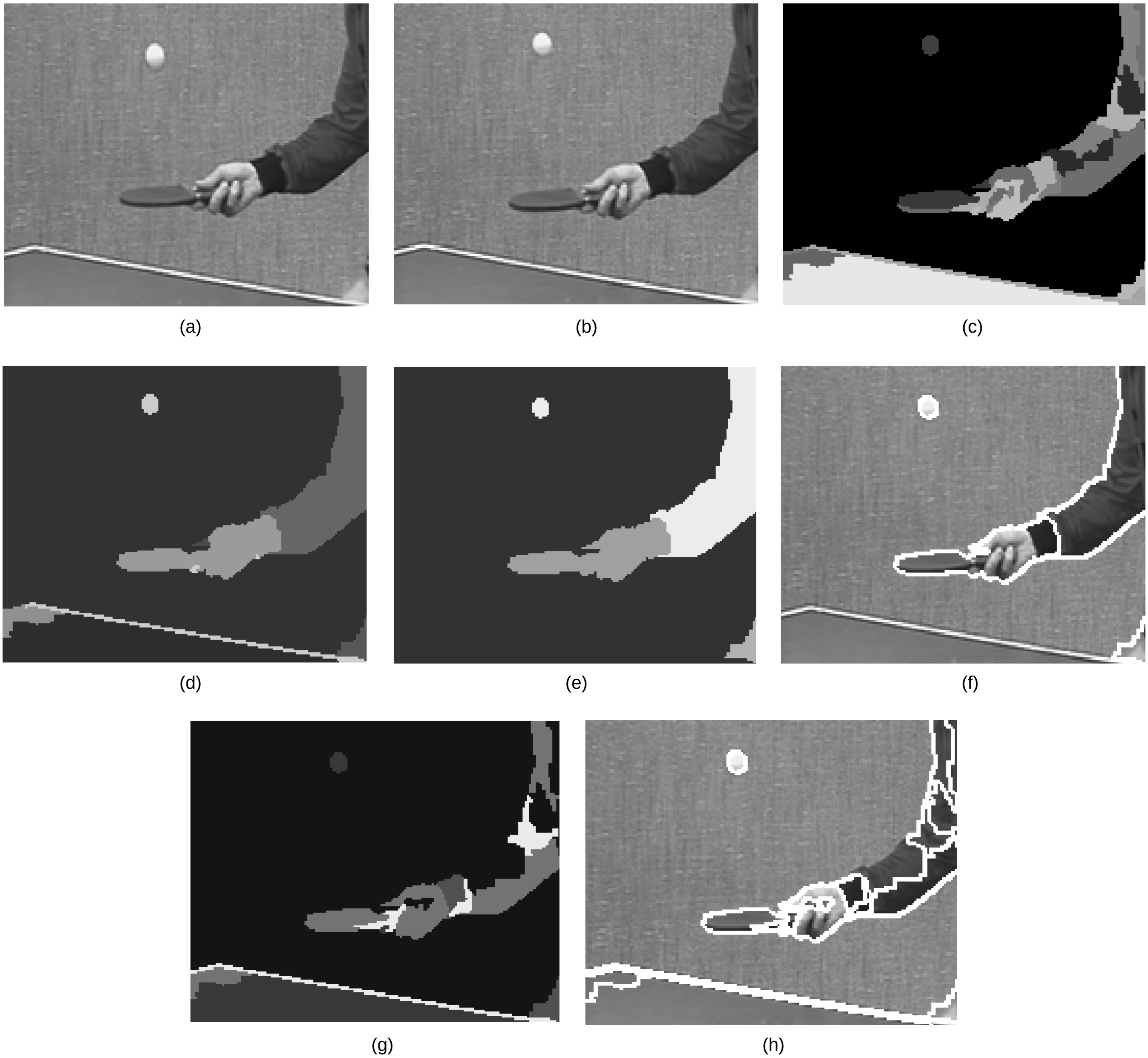


Fig. 9. “Table Tennis”: Spatiotemporal segmentation using arbitrarily shaped initial regions. (a) Previous frame after global motion compensation. (b) Current frame. (c) Set of initial regions to be merged (31 regions). (d) Intermediate spatiotemporal segmentation obtained from the strong rule (11 regions). (e) Final spatiotemporal segmentation obtained after the weak rule (five regions). (f) Boundaries of the final segmentation superimposed onto the current frame. (g) Spatiotemporal segmentation obtained using the technique of Dufaux et al. by requiring a five-region partition. (h) Boundaries obtained by Dufaux’s method superimposed onto the current frame.

information is given a significant weight. Figs. 10a and 10b show the previous frame and the current frame, respectively. Again, the previous frame used here is preprocessed to remove the effects of camera motion, by compensating it for global motion. The set of initial regions shown in Fig. 10c contains 105 regions. After applying the strong rule, 21 regions remain. This intermediate result is shown in Fig. 10d. In this image, the background, the face, and the helmet are clearly discernible. Finally, the weak rule produces a segmentation of the scene reflecting the presence of five objects: the background, the right part of the face and the neck, the left part of the face, the back of the helmet, and the rest of the helmet (see Figs. 10e and 10f). We believe the reason for the relatively poor results lies in the use of the affine motion model. The affine motion model is unable to characterize the motion present in the scene sufficiently well. In this case, a perspective motion model would be expected to produce better results. The spatiotemporal segmentation based on the method proposed by Dufaux et al. (see Figs. 10g and 10h) is obtained by requiring the algorithm to partition the scene into two regions (the head of Foreman and the background). Even with this additional information, the resulting spatiotemporal segmentation does not isolate the head of the man properly. For example, note the hole in the region of the chin and the thin strip of the background region between the face and helmet.

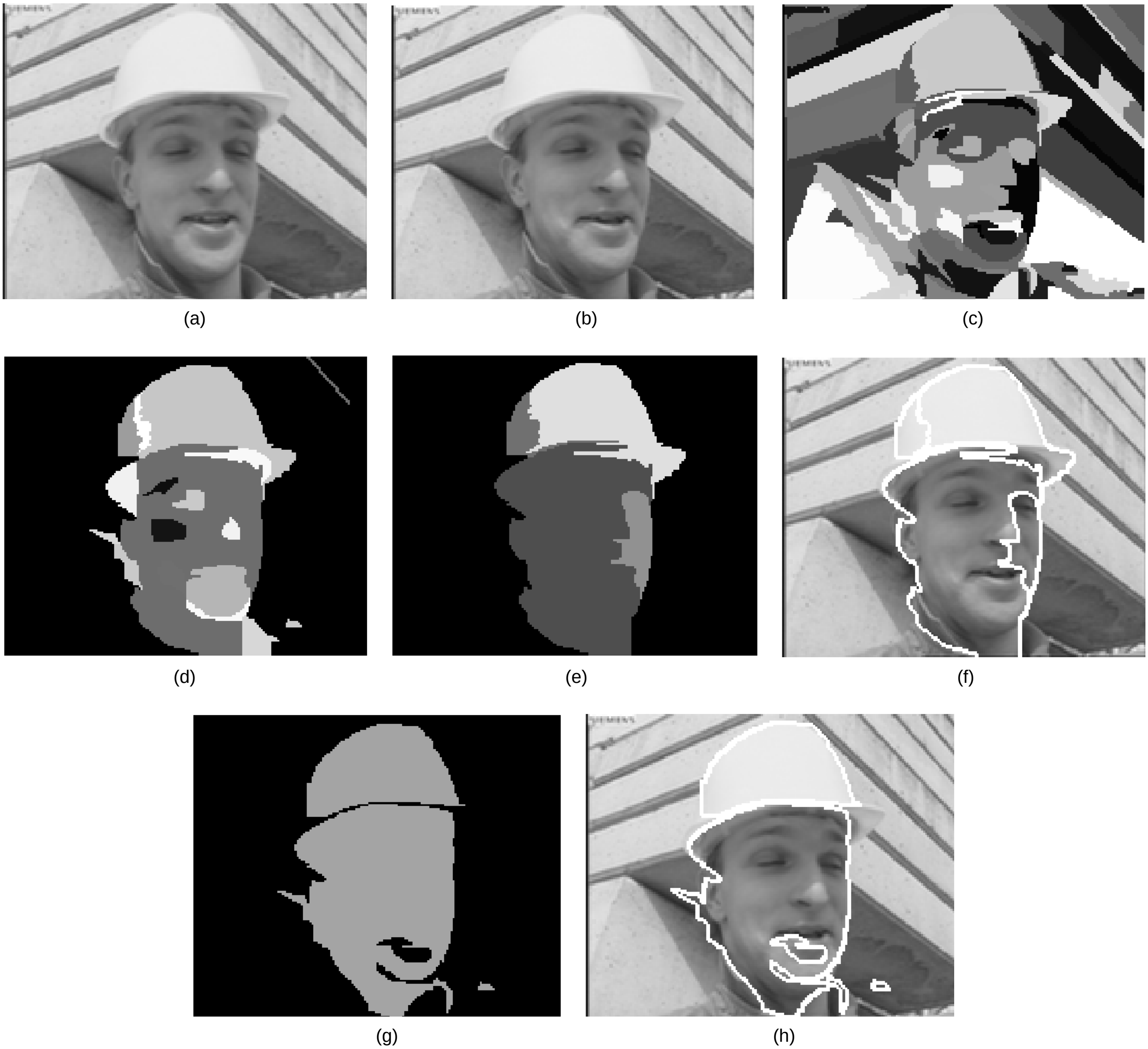


Fig. 10. “Foreman”: Spatiotemporal segmentation using arbitrarily shaped initial regions. (a) Previous frame after global motion compensation. (b) Current frame. (c) Set of initial regions to be merged (105 regions). (d) Intermediate spatiotemporal segmentation obtained from the strong rule (21 regions). (e) Final spatiotemporal segmentation obtained after the weak rule (five regions). (f) Boundaries of the final segmentation superimposed onto the current frame. (g) Spatiotemporal segmentation obtained using the technique of Dufaux et al. by requiring a two-region partition. (h) Boundaries obtained by Dufaux’s method superimposed onto the current frame.

7.2 Effect of Luminance Information

As described in Section 5.3, the proposed spatiotemporal similarity integrates both temporal and spatial information into a single measure. In this section, we demonstrate why spatial information is useful, even though the main criterion for defining objects is temporal coherence.

Fig. 11 shows the different stages of the spatiotemporal segmentation for a frame of the “Table Tennis” sequence, using two different values of the luminance factor (f_L = 1.0, and 0.0, respectively). The case of f_L = 1.0 implies maximum use of the spatial information, while in the other case no spatial information is utilized. In this experiment, the current frame corresponds to the time when the ball is nearly at the top of its trajectory. Consequently, the motion of the ball is very small. When f_L is set to 1.0, the proposed spatiotemporal segmentation technique is able to segment out the ball (see Figs. 11d, 11e, and 11f). This is not possible when no spatial information is used (i.e., f_L = 0.0). (The corresponding results are shown in Figs. 11g, 11h, and 11i.) The motion of the ball is indeed too small to discriminate it from the background.

7.3 Effect of the Pondering Factor λ

In Section 5.2, we stipulate that the temporal similarity should rely on two kinds of temporal information: that embodied in the parametric representation as well as the information present in the residual distribution. Although the temporal information from the residual distribution is

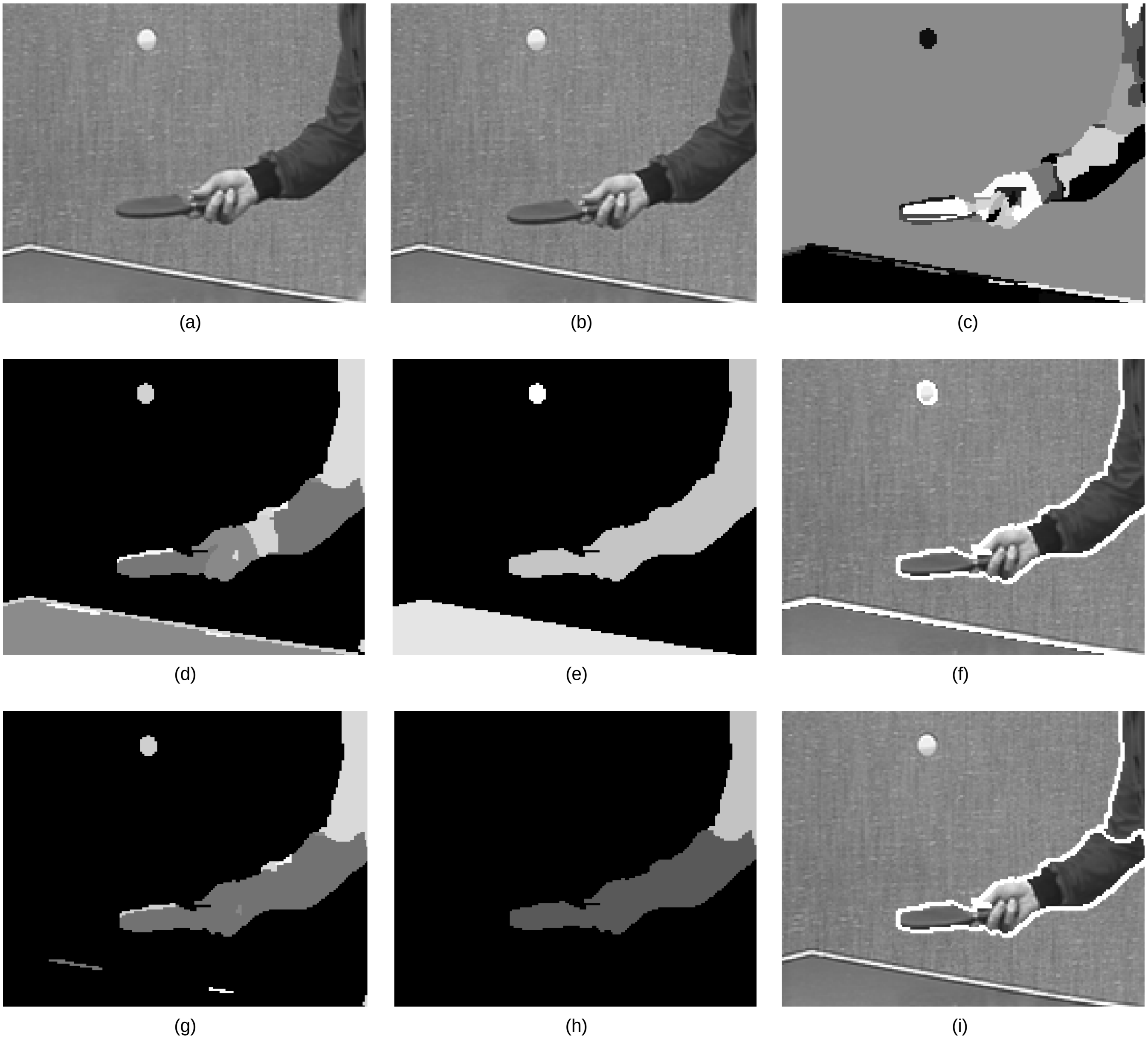


Fig. 11. "Table Tennis": Effect of the luminance information on the spatiotemporal segmentation. (a) Previous frame after global motion compensation. (b) Current frame. (c) Set of initial regions. (d) Spatiotemporal segmentation with luminance factor f_L = 1.0 after applying only the strong rule. (e) Spatiotemporal segmentation with luminance factor f_L = 1.0 after applying both rules. (f) The boundaries of the segments in (e) superimposed on the current frame. (g) Spatiotemporal segmentation with luminance factor f_L = 0.0 after applying only the strong rules. (h) Spatiotemporal segmentation with f_L = 0.0 after applying both rules. (i) The boundaries of the segments in (h) superimposed on the current frame.

given more importance, the parametric information is also quite useful. In this experiment we evaluate the contribution of the parametric information, which is regulated by the pondering factor, λ.

The importance of the pondering factor is demonstrated by slightly modifying the conditions of the experiments depicted in Fig. 8. Suppose we do not use the pondering factor (i.e., λ = 1.0). In this case, measure of temporal similarity draws on only the residual distribution. The corresponding results are presented in Fig. 12. The spatiotemporal segmentation identifies three objects: the background, the head with the right part of the body, and the left part of the body. Although only the temporal information available in the residual distribution is used, the resulting decomposition of the scene is fairly accurate. However, the additional contribution of the parametric information would have allowed a better definition of the objects. This is apparent by comparing Fig. 8f with Fig. 12d.

8 Summary and Conclusions

In this paper, a technique for unsupervised spatiotemporal segmentation has been proposed. Only the information present in two consecutive frames is used. The proposed technique takes a set of initial regions as additional input. This set typically represents an oversegmentation, resulting from some other segmentation algorithm. In this work, we have used the segmentation algorithm described in Section 4

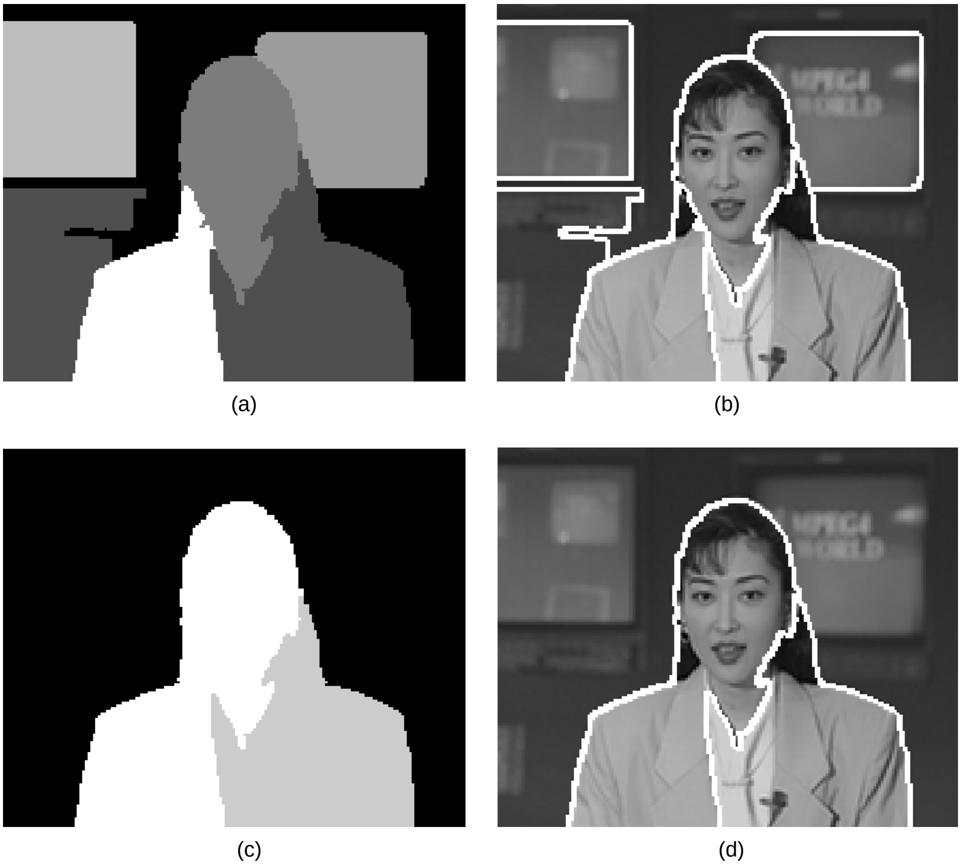


Fig. 12. “Akiyo”: Effect of not using the pondering factor λ. (a) Spatiotemporal segmentation after applying only the strong rule. (b) The boundaries of the regions in (a), superimposed on the current frame. (c) Final spatiotemporal segmentation. (d) The boundaries of the regions in (c), superimposed on the current frame. Here, the previous frame and current frame are the same as shown in Fig. 8a and 8b, respectively.

to generate the set of initial regions. However, the proposed region-merging algorithm can be used with other segmentation techniques as well. Results using a quadtree-based initial segmentation are presented in [40].

The regions are iteratively merged in order to determine the objects forming the scene. The regions are merged on the basis of their mutual spatiotemporal similarities. Defined as a combination of temporal and spatial information, the spatiotemporal similarity is formulated in the statistical framework as a hypothesis test. To that end, a new test statistic for the temporal information, the MKS test, is presented. It permits the simultaneous use of the temporal information available in the residual distribution and in the motion parametric representation. The merging of the regions is carried out by using a weighted, directed graph. Two graph-based clustering rules are proposed. They are called the strong rule and the weak rule. The strong rule is applied first, followed by the weak rule. Each rule is applied iteratively. The graph is updated at the end of each iteration. The proposed graph-based clustering strategy takes into account the specificities of the problem at hand. It is able to efficiently exploit the information represented in the graph, while being robust to erroneous motion information as well as to oddly defined regions.

The proposed technique has been evaluated on different types of video sequences. The spatial and temporal precision of the object boundaries depends on the quality of the initial regions. Our experiments also demonstrate the importance of using spatial information. Such information may contribute significantly when temporal information is not discriminatory enough. Finally, the necessity of combining the temporal information existing in both the residual distribution and the parametric representation has been demonstrated.

An extension of the technique presented here to recursive spatiotemporal segmentation and an unsupervised object-tracking algorithm is proposed in [40]. It is a recursive procedure for spatiotemporal segmentation through the sequence. It combines the tracking information into the segmentation process. The spatiotemporal segmentation of the successive frames is robustly derived, and the objects forming the scene are tracked through time.

REFERENCES

[1] P. Bouthemy and E. Francois, “Motion Segmentation and Qualitative Dynamic Scene Analysis From an Image Sequence,” *Int’l J. Computer Vision*, vol. 10, no. 2, pp. 157-182, 1993.

[2] G. Adiv, “Determining Three-Dimensional Motion and Structure From Optical Flow Generated by Several Moving Objects,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 7, no. 4, pp. 384-401, July 1985.

[3] F. Dufaux, F. Moscheni, and A. Lippman, “Spatiotemporal Segmentation Based on Motion and Static Segmentation,” *Proc. ICIP’95*, vol. 1, pp. 306–309, Washington, DC, Oct. 1995.

[4] M. Kunt, A. Ikonomopoulos, and M. Kocher, "Second Generation Image Coding Techniques," *Proc. IEEE*, vol. 73, no. 4, pp. 549-575, Apr. 1985.

[5] F. Dufaux and F. Moscheni, "Background Mosaicking for Low Bit Rate Video Coding," *Proc. ICIP'96*, vol. I, pp. 673-676, Lausanne, Switzerland, Sept. 1996.

[6] M. Allmen and C.R. Dyer, "Computing Spatiotemporal Relations for Dynamic Perceptual Organisation," *CVGIP: Image Understanding*, vol. 58, no. 3, pp. 338-351, Nov. 1993.

[7] S. Ayer, P. Schroeter, and J. Bigün, "Segmentation of Moving Objects by Robust Motion Parameter Estimation Over Multiple Frames," *ECCV'94*, vol. 2, pp. 316-327, Stockholm, Sweden, May 1994.

[8] J.L. Barron, D.J. Fleet, and S.S. Beauchemin, "Performance of Optical Flow Techniques. *Int'l J. Computer Vision*, vol. 12, no. 1, pp. 43-77, 1994.

[9] J.R. Bergen, J.B. Burt, J. Hingorani, and S. Peleg, "A Three-Frame Algorithm for Estimating Two-Component Image Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, pp. 886-896, Sept. 1992.

[10] A.D. Jepson and M.J. Black, "Mixture Models for Optical Flow Computation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 760-761, New-York, June 1993.

[11] M. Irani, B. Rousso, and S. Peleg, "Computing Occluding and Transparent Motions," *Int'l J. Computer Vision*, vol. 12, no. 1, pp. 5-16, 1994.

[12] M.J. Black and P. Anandan, "A Framework for the Robust Estimation of Optical Flow," *ICCV'94*, pp. 231-236, Berlin, Germany, May 1993.

[13] H.G. Musmann, M. Hoetter, and J. Ostermann, "Object-Oriented Analysis-Synthesis Coding of Moving Images," *Signal Processing: Image Communication*, vol. 1, no. 2, pp. 117-138, Oct. 1989.

[14] S.F. Wu and J. Kittler, "A Gradient-Based Method for General Motion Estimation and Segmentation," *J. Visual Comm. and Image Representation*, vol. 4, no. 1, pp. 25-38, Mar. 1993.

[15] P.J. Burt, R. Hingorani, and R.J. Kolczynski, "Mechanisms for Isolating Component Patterns in the Sequential Analysis of Multiple Motion," *IEEE Workshop on Visual Motion*, pp. 187-193, Princeton, NJ, 7-9 Oct. 1991.

[16] P.J. Burt, J.R. Bergen, R. Hingorani, R. Kolczynski, W.A. Lee, A. Leung, J. Lubin, and H. Shvaytser, "Object Tracking With a Moving Camera, an Application of Dynamic Motion Analysis," *IEEE Proc. Workshop on Visual Motion*, pp. 2-12, Irvine, Calif., Mar. 1989.

[17] M. Irani, B. Rousso, and S. Peleg, "Detecting and Tracking Multiple Moving Objects Using Temporal Integration," G. Sandini, ed., *Second European Conf. Computer Vision*, pp. 282-287, S.Margherita, Italy, 1992. Springer-Verlag.

[18] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*. New York, NY: John Wiley & Sons, Inc, 1987.

[19] S. Ayer and P. Schroeter, "Hierarchical Robust Motion Estimation for Segmentation of Moving Objects," *IEEE Workshop on Image and Multidimensional Signal Processing*, pp. 122-123, Cannes, France, Sept. 1993.

[20] P. Schroeter and S. Ayer, "Multi-Frame Based Segmentation of Moving Objects by Combining Luminance and Motion," *Proc. EUSIPCO 94*, Edinburgh, U.K., Sept. 1994.

[21] S. Ayer, P. Schroeter, and J. Bigün, "Tracking Based on Hierachical Multiple Motion Estimation and Robust Regression," *Time Varying Image Processing and Moving Objects Recognitions*, Florence, Italy, June 1993.

[22] B. Duc, P. Schroeter, and J. Bigün, "Spatiotemporal Robust Motion Estimation and Segmentation," *Sixth Int'l Conf. Computer Analysis of Images and Patterns*, pp. 238-245, Prague, 6-8 Sept. 1995.

[23] C. Gu, "Multivalued Morphology and Segmentation-Based Coding," PhD thesis, Swiss Fed. Inst. of Technology, Lausanne, 1996.

[24] S. Ayer and H.S. Sawhney, "Layered Representation of Motion Video Using Robust Maximum-Likelihood Estimation of Mixture Models and MDL Encoding," *Fifth Int'l Conf. Computer Vision*, pp. 777-784, Cambridge, Mass., June 1995.

[25] H. Zheng and S.D. Blostein, "Motion-Based Object Segmentation and Estimation Using the MDL Principle," *IEEE Trans. Image Processing*, vol. 4, no. 9, pp. 1,223-1,235, Sept. 1995.

[26] R. Szeliski and H. Shum, "Motion Estimation With Quadtree Splines," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 12, pp. 1,199-1,211, Dec. 1996.

[27] P. Salembier, L. Torres, F. Meyer, and C. Gu, Region-Based Video Coding Using Mathematical Morpholgy," *Proc. IEEE*, vol. 83, no. 6, pp. 843-857, June 1995.

[28] P. Anandan, J.R. Bergen, K.J. Hanna, and R. Hingorani, "Hierarchical Model-Based Motion Estimation," M.I. Sezan and R.L. Lagendijk, eds., *Motion Analysis and Image Sequence Processing*, pp. 1–22. Kluwer Academic Publishers, 1993.

[29] J.Y.A. Wang and E.H. Adelson, "Spatiotemporal Segmentation of Video Data," *SPIE Proc. Image and Video Processing II*, vol. 2,182, San Jose, Calif., Feb. 1994.

[30] G. Adiv, "Inherent Ambiguities in Recovering 3D Motion and Structure From a Noisy Flow Field," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 5, pp. 477-489, May 1989.

[31] D.W. Murray and B.F. Buxton, "Scene Segmentation From Visual Motion Using Global Optimization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, no. 2, pp. 220-228, Mar. 1987.

[32] P. Bouthemy and J. Santillana Rivero, "A Hierarchical Likelihood Approach for Region Segmentation According to Motion-Based Criteria," *ICCV'87*, pp. 463-467, London, UK, 1987.

[33] F. Dufaux and F. Moscheni, "Segmentation-Based Motion Estimation for Second Generation Video Coding Techniques," L. Torres and M. Kunt, eds., *Video Coding: The Second Generation Approach*, pp. 219-263. Kluwer Academic Publishers, 1995.

[34] M.J. Black, "Combining Intensity and Motion for Incremental Segmentation and Tracking Over Long Image Sequences," G. Sandini, ed., *Second European Conf. Computer Vision*, pp. 485-493, Santa Margherita, Italy, May 1992.

[35] F. Heitz and P. Bouthemy, "Multimodal Estimation of Discontinuous Optical Flow Using Markov Random Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 12, pp. 1,217-1,232, Dec. 1993.

[36] H. Gu, Y. Shirai, and M. Asada, "MDL-Based Segmentation and Motion Modelling in a Long Image Sequence of Scene With Multiple Independently Moving Objects," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 58-64, Jan. 1996.

[37] W.B. Thompson, "Combining Motion and Contrast for Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 2, no. 6, pp. 543-549, Nov. 1980.

[38] B.A. Wandell, *Foundations of Vision*. Sinauer Associates, Inc., 1995.

[39] J.Y.A. Wang and E.H. Adelson, "Representing Moving Images With Layers," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 625-638, Sept. 1994.

[40] F. Moscheni, "Spatiotemporal Segmentation and Object Tracking: An Application to Second Generation Video Coding," PhD thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, 1997.

[41] W.B. Thompson and T.-G. Pong, "Detecting Moving Objects," *Int'l J. Computer Vision*, vol. 4, pp. 39-57, 1990.

[42] J.W. Park and S.U Lee, "Joint Image Segmentation and Motion Estimation for Low Bit Rate Video Coding," *Proc. ICIP'96*, vol. 2, pp. 501–504, Lausanne, Switzerland, Sept. 1996.

[43] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York, NY: McGraw-Hill, 1991.

[44] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C*. Cambridge, Mass.: Cambridge University Press, 1992.

[45] P. Meer, D. Mintz, A. Rosenfeld, and D.Y. Kim, "Robust Regression Methods for Computer Vision: A Review," *Int'l J. Computer Vision*, vol. 6, no. 1, pp. 59-70, 1991.

[46] A.K. Jain and C.D. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[47] M. Schütz and T. Ebrahimi, "Matching Error Based Criterion of Region Merging for Joint Motion Estimation and Segmentation Techniques," *Proc. ICIP'96*, vol. II, pp. 509-512, Lausanne, Switzerland, Sept. 1996.

[48] F. Moscheni, F. Dufaux, and M. Kunt, "A New Two-Stage Global/Local Motion Estimation Based on a Background/Foreground Segmentation," *Proc. ICASSP'95*, vol. 4, pp. 2,261-2,264, Detroit, MI, May 1995.

[49] L. Kaufman and P.J. Rousseeuw,. *Finding Groups in Data*. New York, NY: John Wiley&Sons, Inc., 1990.

Fabrice Moscheni received an MS in physics from the Swiss Federal Institute of Technology at Lausanne (EPFL). He worked at Fujitsu Laboratories Ltd., Atsugi, Japan and then joined the Signal Processing Laboratory of EPFL as a research assistant and, later, as a PhD student. He received his PhD in Communications Technology in 1997. His work involved the supervision of EPFL and ERASMUS students as well as taking part in the European projects dTTb and Vadis. Since the beginning of 1998, he has been working for FASTCOM Technology, in Lausanne. He is in charge of developing software applications for FASTCOM's video surveillance products, which are stand-alone embedded machine vision tools for networking usages. His research interests are in video sequence processing and analysis, motion estimation, second generation coding, and higher-order statistics. He is a member of the Swiss Technical Society and IEEE.

Sushil Bhattacharjee received his MS in computer science from Michigan State University, East Lansing, Michigan. He is pursuing a PhD degree in electrical engineering at the Signal Processing Lab of the Swiss Federal Institute of Technology, Lausanne (EPFL). His professional background includes image-processing and pattern recognition. He has worked in the areas of geometric correction of remote-sensing imagery, texture-analysis based on multichannel filtering, document-image analysis, and cartographic image interpretation. His current research interests are in content-based retrieval of images, spatiotemporal segmentation of video, and video-coding. He has coauthored published works on all of these topics.

Muran Kunt (S'69-M'74-SM'78-F'86) received his MS in physics and his PhD in electrical engineering, both from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 1969 and 1974, respectively. From 1974 to 1976, he was a visiting scientist at the Research Laboratory of Electronics of the Massachusetts Institute of Technology, where he developed compression techniques for X-ray images and electronic image files. In 1976, he returned to the Swiss Federal Institute of Technology (EPFL) where, presently, he is professor of electrical engineering and director of the Signal Processing Laboratory, one of the largest laboratories at EPFL. He conducts teaching and research in digital signal and image processing with applications to modeling, coding, pattern recognition, scene analysis, industrial developments, and biomedical engineering. His laboratory participates in a large number of European projects under various programs such as Esprit, Eureka, Race, HCM, Commett, and Cost. He is the author or the coauthor of more than two hundred research papers and fourteen books and hold seven patents. He is the editor-in-chief of the Signal Processing Journal and a founding member of EURASIP, the European Association for Signal Processing. He serves as a chairman and/or a member of the scientific committees of several international conferences and in the editorial boards of the Proceedings of the IEEE, Pattern Recognition Letters and Traitement du Signal. He was the cochairman of the First European Signal Processing Conference, which was held in Lausanne in 1980 and the general chairman of the International Image Processing Conference (ICIP'96) held in Lausanne in 1996. He was the president of the Swiss Association for Pattern Recognition from its creation until 1997. He consults for governmental offices including the French General Assembly. He received the gold medal of EURASIP for meritorious services and the IEEE ASSP technical achievement award in 1983 and 1997, respectively.