ROBUST REGION MERGING FOR SPATIO-TEMPORAL SEGMENTATION

Fabrice Moscheni *Sushil Bhattacharjee*

Signal Processing Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
{moscheni,sushil}@ltssg4.epfl.ch

ABSTRACT

A region merging technique for spatio-temporal segmentation of scenes is presented here. The proposed technique is a bottom-up method and expects an initial set of regions. These regions are compared on the basis of a similarity measure that integrates both spatial and temporal information. The unsupervised merging procedure is based on a weighted, directed graph that is updated dynamically. Two graph based clustering rules are presented. These rules are used to cluster regions into ensembles that represent meaningful objects present in the scene. Experimental results demonstrate the efficiency of the proposed method.

1. INTRODUCTION

This paper addresses the problem of extracting semantically meaningful objects within the framework of spatio-temporal segmentation. Spatio-temporal segmentation attempts to describe a dynamic scene in terms of moving objects. Applications like object tracking [1] and structure from motion can benefit from such a visually meaningful segmentation.

The techniques for spatio-temporal segmentation can be grouped into two broad classes: the top-down approach and the bottom-up approach. Top-down approaches iteratively estimate the parameters of dominant motions in the scene. Regions complying with the current dominant motion are assumed to belong to the same object, and are not considered in the next iteration [2]. In contrast, region merging (bottom-up) approaches start with a set of regions and merge them into moving objects according to some spatio-temporal criteria. Several criteria for region-merging have been proposed in the relevant literature. However, in general, these methods fail to exploit the available information in its entirety. In this paper, we propose a region merging algorithm for spatio-temporal segmentation that attempts to overcome the drawbacks of other previously proposed algorithms.

The paper is structured as follows. In Sec. 2 we present a brief review of previous relevant works. The proposed technique for spatio-temporal segmentation is presented in Sec. 3. Section 4 provides experimental results to demonstrate the robustness of the proposed approach. Finally, conclusions are drawn in Sec. 5.

2. BACKGROUND

Bottom-up spatio-temporal segmentation methods assume that the frame of interest in the image sequence has already been segmented into a set of regions. The manner in which these regions are obtained may vary. They may be chosen arbitrarily (e.g., blocks) or they may be the result of some segmentation based on spatial/temporal information. For instance, the techniques proposed by Adiv [3] and Wang and Adelson [4] rely only on motion information. They define regions based on their consistency with a specified motion model. Dufaux *et al.* [5] use both spatial and temporal information. In the limiting case, each pixel can be considered as a separate region.

The region merging process has two important components: the definition of the region-similarity measure, and its usage. Several similarity measures have been proposed for spatio-temporal region merging. They differ in the way they use (or do not use) the spatial information and the temporal information given by the motion parameters. The second issue is the way in which the similarity measure is used to determine whether regions should be merged.

Some region merging techniques for spatio-temporal segmentation use temporal information available only on the motion parameter space. The methods of Dufaux *et al.* [5], and Wang and Adelson [4] define the region-similarity measures to be the distances in the motion parameter space. The merging decision is based on a clustering procedure and regions assigned to the same cluster are merged into a single moving object. This method is obviously sensitive to error in motion estimation and to the distance measure used in the clustering process. Also, the number of clusters (i.e., objects) has to be predetermined by the user or computed by some *ad hoc* method. A more severe problem is the following. Depending on the scene and the motion model chosen, similar optical flows may be represented by very different sets of motion parameters [6]. In other words, the parameterization of the motion need not have a unique solution and the hypothesis that the motion parameters represent the entire motion information may indeed be wrong. This implies that two regions which are moving in a similar way may turn out to have very different motion parameters and, thus, will not be merged by the clustering procedure.

Other regions merging methods for spatio-temporal segmentation attempt to use all available information. Both spatial and temporal information are exploited. For in-

stance, the similarity measure proposed by Adiv [3] relies on the distribution of the residuals obtained after motion compensation. In this approach, spatial information is incorporated by restricting region merging to adjacent regions only. The merging decision is performed based on the variance of the residual distributions. Two regions are merged if the variance of the newly formed region is similar to those of the individual regions before merging. However, the use of a single statistic is too drastic a reduction of dimensionality. It may induce wrong merging decisions, since all the information present in the distribution is not examined. Furthermore, no use is made of the motion information available in parametric form.

Moscheni and Dufaux have also proposed an approach to spatio-temporal segmentation based on region merging where the region-similarity measure utilizes both spatial and temporal information[7]. Temporal information is exploited in both its residual-distribution form and its parametric representation. Spatial information is used to ensure that only adjacent regions are merged. The spatio-temporal similarities among the various regions are represented in a weighted, directed graph. The algorithm clusters the vertices of the graph, by applying two clustering rules sequentially. However, the similarity measure used in this algorithm exploits the available spatial information poorly. It simply imposes that only adjacent regions may be merged based on temporal information. The clustering procedure also suffers from some drawbacks. First, the weights for the edges in the graph are binary-valued. A better use of the similarity measure would be to use it in its continuous-valued form. Secondly, as the vertices of the graph get merged to form new vertices, the edges in the graph are not updated dynamically.

3. PROPOSED SPATIO-TEMPORAL REGION MERGING

In this paper we propose an algorithm for spatio-temporal region merging. This unsupervised procedure is robust in the presence of outliers. Both spatial and temporal information are used to guide the merging process which is carried out based on a dynamic clustering of regions. The algorithm assumes that an initial segmentation of the scene, and the motion parameters for each region, are available. Any method may be used to generate the initial segmentation. The patio-temporal similarity measure and the dynamic graph clustering strategy are discussed in this section.

3.1. Spatio-temporal Similarity Measure

The region-similarity measure proposed here exploits both spatial and temporal information. However, we place more emphasis on temporal information as we are looking for coherently moving objects. The spatio-temporal similarity F_{AB}, between two regions, **A** and **B**, is defined as a combination of the result, T_{AB}, of a test statistic on the temporal information and the result, S_{AB}, of a test statistic on the spatial information. These are discussed below.

Test Statistic for Temporal Information: The determination of the temporal information is based on a specific motion model. We use a fully parametric affine model in this work. The test statistic on the temporal information determines the level of coherence between the motions of the two regions **A** and **B**. We use the Modified Kolmogoroff-Smirnov test (MKS test) [7] for this purpose. This non-parametric test statistic exploits motion information available in the motion parameters as well as that present in the residual obtained after compensating region **A** using the motion parameters of region **B**. Using the MKS test, we define the significance level, T_{AB}, that region **A** moves in the same way as region **B**. Spatial information is also used at this stage as T_{AB} is computed only for adjacent regions, and is set to 0 otherwise.

Test Statistic for Spatial Information: In the context of spatio-temporal region merging, the spatial information is used to determine the probability that two regions share spatial characteristics. Spatial information may range from shape information to texture information. In this paper, the spatial similarity of two adjacent regions **A** and **B** is computed based on the respective medians, L_{AB} and L_{BA}, of their luminance values along their common border. The hypothesis that the regions **A** and **B** are spatially similar is given by the significance level S_{AB} defined by:

$$S_{AB} = Prob(|q| \geq |L_{AB} - L_{BA}|), \quad (1)$$

where q is the maximum likelihood estimator of the hypothesis that L_{AB} and L_{BA} are equal. Assuming the set of luminance medians to be trials of a Gaussian variable with variance σ, it can be shown that q is a random variable with $q \sim N(0, \sqrt{2}\sigma)$. In practice the variance σ is estimated over the ensemble of luminance medians of all the regions. If the regions **A** and **B** are not adjacent, S_{AB} is set to 0.

The Spatio-Temporal Similarity Measure: The spatio-temporal similarity, F_{AB}, of two regions **A** and **B** is specified as a combination of the significance levels T_{AB} and S_{AB}. Recall, however, that spatio-temporal region merging is performed to obtain objects that are likely to be composed of regions having different spatial characteristics. Thus, F_{AB} must rely mainly on the temporal information (i.e. T_{AB}). Let Υ_X be the ensemble of neighboring regions for region **X**. The proposed spatio-temporal similarity measure F_{AB} is written as follows:

$$F_{AB} = T_{AB} - k\, T_{AB}\, (M - S_{AB}), \quad (2)$$

with

$$M = [\max(\max(S_{AJ}), \max(S_{RI})) \mid \mathbf{J} \in \Upsilon_A,\ \mathbf{I} \in \Upsilon_R],$$
$$\mathbf{R} = (\mathbf{J} | S_{AJ} \text{ is maximum}) \text{ and } k \in [0, 1].$$

Equation 2 reflects the fact that T_{AB} is the most significant term in the spatio-temporal similarity F_{AB}. S_{AB} is just used as a corrective factor. The factor M conveys information about the general spatial coherence of region **A** with its neighboring regions. The factor k has a preset value and controls the level of the correction based on spatial coherence information.

3.2. Graph-Based Dynamic Clustering Strategy

The region merging procedure is developed in the framework of graph theory [7], and uses the spatio-temporal similarity measure presented in Sec. 3.1. More precisely, the spatio-temporal similarity measure is used to construct a

graph where the vertices represent the regions and edges represent the similarity between regions. The weights of the edges are expressed as percentages. Clearly, this graph is weighted as well as directed. Regions are merged according to the information represented in this graph. The clustering strategy employs two rules, referred to as *the strong rule* and *the weak rule*, respectively. These rules are explained below.

In the following, the set $\mathcal{F}$ is the set of indices of the F regions R_f to be clustered. That is, $\mathcal{F} = (1, \cdots, F)$. In the same way, $\mathcal{I}$ is the set of indices of I clusters, C_i. Therefore, $\mathcal{I} = \{1, \cdots, I\}$. The ensemble $\mathcal{I}$ is by definition a subset of the ensemble $\mathcal{F}$. The strong rule can now be defined as follows:

$$C_i = \{R_m, m \in \mathcal{F} \mid \exists (R_l \text{ and } R_k, k, l \in \mathcal{F}; R_k, R_l \in C_i) \text{ such as } (R_m \rightarrow R_l \text{ and } R_k \rightarrow R_m)\}, \; i \in \mathcal{I}. \quad (3)$$

Here, $B \rightarrow A$ denotes that the spatio-temporal similarity of region **A** with region **B** is greater than the current strong rule threshold t_{sr}. The strong will thus merge regions **A** and **B** only if F_{AB} and F_{BA} are greater than t_{sr}.

The weak rule aims at relaxing the conditions for merging regions imposed by the strong rule. Further, when the weak rule is applied, the graph is not updated every time two regions are merged. This type of merging is referred to as non-dynamic merging. For each cluster C_j, initially containing exactly one region, the weak rule first determines the ensemble Ω of clusters C_i with which C_j could be merged. The ensemble Ω is defined by:

$$\Omega = \left\{ C_i, i \neq j \;\middle|\; \sum_{R_k \in C_i} \sum_{R_l \in C_j} (R_k \rightarrow R_l) \geq \text{Card}(C_j) \right\}, \quad (4)$$

where $\text{Card}(C_j)$ denotes the cardinality of the cluster C_j. $B \rightarrow A$ denotes that the spatio-temporal similarity of region **A** with region **B** is greater than the current threshold for the weak rule.

Assuming that Ω is not empty, an important step in the non-dynamic merging consists in selecting the cluster $C_s \in \Omega$, with which the cluster C_j has to be merged. This selection is performed as follows:

$$C_s = \{C_i \in \Omega \text{ , such as } \max_{\mu} \max_{\pi} \max_{\chi}\} , \quad (5)$$

where

$$\chi = \sum_{R_k \in C_i} \sum_{R_l \in C_j} (R_k \rightarrow R_l) ,$$

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

$$\mu = \sum_{R_k \in C_i} \sum_{R_l \notin C_i} (R_k \rightarrow R_l) ,$$

where $Area(C_i)$ is the area of the cluster C_i. The non-dynamic merging is carried out by iteratively applying Eq. 5 on the set of initial regions while the current weak rule threshold decreases from 100% by steps of 1% to its lowest allowed value. Thus, the weak rule merging permits a robust hierarchical region merging that does not require any update of the graph.

The strong and weak rules are applied successively in the context of a dynamic graph updating strategy. First, the regions represented in the graph iteratively clustered using only the strong rule. After each iteration, the graph representing the relationships among regions 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. Initially set of 100%, the threshold value for the strong rule, t_{sr}, is recomputed after each iteration. The maximum value E_s that would still allow the strong rule to carry out a merging, is determined from the graph. The threshold t_{sr} is thus defined as:

$$t_{sr} = 100 - Q_{sr} \; (1 + (Int)((100 - E_s)/Q_{sr})), \quad (6)$$

where Q_{sr} is a pre-defined step-size for lowering t_{sr}. The iteration stops when t_{sr} is less than the pre-defined lowest threshold threshold t_{lsr}. The weak rule is now applied iteratively to the remaining regions. Again, the graph is updated after each iteration step, while the non-dynamic merging described above occurs within each iteration. Similar to t_{sr}, the threshold t_{wr} is lowered at every iteration as follows:

$$t_{wr} = 100 - Q_{wr} \; (1 + (Int)((100 - E_w)/Q_{wr})), \quad (7)$$

where Q_{wr} is a pre-defined step-size for lowering t_{wr}, and E_w is the maximum value in the graph that would still allow the weak rule to carry out a merging. The merging process stops when t_{wr} is lower than the pre-defined lowest threshold t_{lwr}.

4. EXPERIMENTAL RESULTS

Two sets of results are presented for the proposed region merging technique. In the first set, the region merging is based on an initial set of regions obtained using a method similar to the one used by Dufaux *et al.* [5]. We also present region merging results for the case when the initial set of regions is obtained using a quadtree based segmentation. This demonstrates that the proposed region merging method can be used with different methods of generating the initial set of regions. Experiments have been carried out on successive frames of sequences in QCIF format.

Figure 1(a) shows one frame of the 'Table Tennis' sequence. As mentioned before, the proposed algorithm expects an initial set of regions as input. Figure 1(b) shows the initial set of regions of the image in Fig. 1(a). This set of regions is obtained by clustering pixels in the three-dimensional color space in combination with a motion based refinement as proposed in [5]. The proposed method merges these regions to form meaningful objects. Figure 1(c) shows the segmentation produced by applying the strong rule only. This segmentation is further improved by applying the weak rule. The final segmentation result is shown in Fig. 1(d). Five objects, namely, the ball, the table, the arm, the racquet and the hand holding it, and the background, are obtained. Note that the cuff of the shirt and part of the forearm have been clustered together with the hand and racquet. This is explained by the fact that in this section of the sequence, these regions have very coherent motion.

Figure 2 shows the spatio-temporal segmentation results when the initial set of regions for Fig. 1(a) is obtained using a quadtree based segmentation approach. Initially the input frame is divided into arbitrary blocks, and their motion parameters are estimated. Blocks that show a *displaced frame difference* (DFD) higher than a preset threshold are split into four equal sized blocks. This process is repeated until a lower limit on the size of blocks is reached. Figure 2(a) shows the obtained initial set of regions. The result of merging these regions after applying the strong rule only is shown in Fig. 2(b). We note that the final result, shown in Fig. 2(c), is quite similar to the result shown in Fig. 1(d). The ball, and the arm and racquet (as one region) are identified quite well. Of course, the regions appear blocky, because the region merging algorithm starts with an initial set of blocky regions. Note that, in this experiment, the region merging process fails to identify the table.

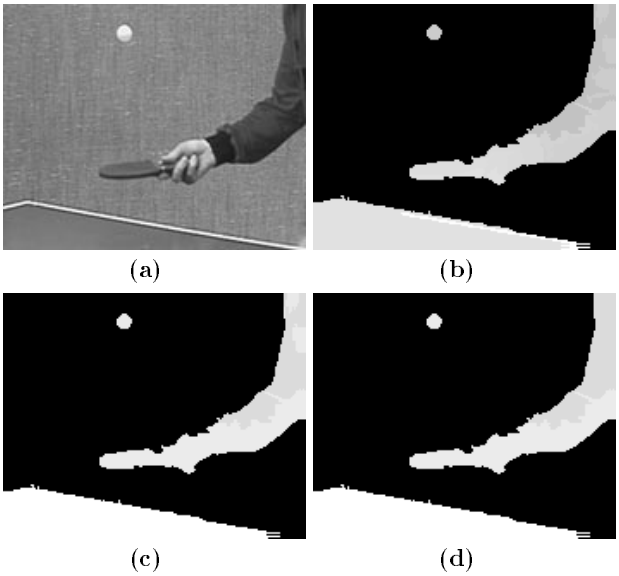

Figure 1: "Table Tennis": **(a)** one frame, **(b)** initial set of regions (20 regions); **(c)** spatio-temporal segmentation obtained after applying the strong rule (7 regions); **(d)** final spatio-temporal segmentation (5 regions).

5. CONCLUSION

We have presented an unsupervised region-merging technique for spatio-temporal segmentation. The proposed measure of region-similarity efficiently exploits both temporal and spatial information. The merging process is based on a graph which is used to represent the spatio-temporal coherence of regions. Two clustering rules are successively applied to this graph, in order to merge regions. The graph is dynamically updated during the region merging process.

Experimental results demonstrate the effectiveness of the proposed method, and show that the region merging process works well with different sets of regions.

6. REFERENCES

[1] F. Moscheni, F. Dufaux, and M. Kunt. Object tracking based on temporal and spatial information. In *IEEE Proc. ICASSP'96*, Atlanta, GA, May 1996.

[2] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In Sandini G., editor, *Second European Conference on Computer Vision*, pages 282–287. Springer-Verlag, 1992.

[3] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. PAMI-7, no. 4, pp. 384-401, July 1985.

[4] J.Y.A. Wang and E.H. Adelson. Spatio-temporal segmentation of video data. In *SPIE Proc. Image and Video Processing II*, volume 2182, San Jose, CA, February 1994.

[5] F. Dufaux, F. Moscheni, and A. Lippman. Spatio-temporal segmentation based on motion and static segmentation. In *IEEE Proc. ICIP'95*, volume 1, pages 306–309, Washington, DC, October 1995.

[6] G. Adiv. Inherent ambiguities in recovering 3d motion and structure from a noisy flow field. *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-11, no. 5, pp. 477-489, May 1989.

[7] F. Moscheni and F. Dufaux. Region merging based on robust statistical testing. In *SPIE Proc. Visual Communications and Image Processing '96*, Orlando, Florida, March 1996.

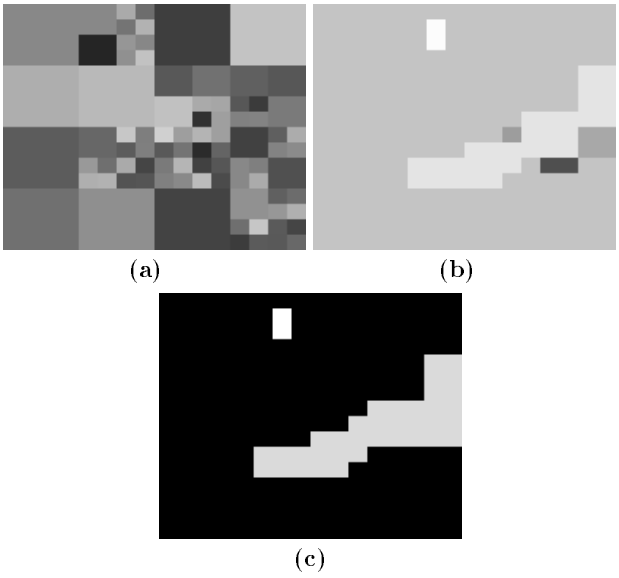

Figure 2: **(a)** initial set of regions (85 blocks) obtained using quadtree based segmentation; **(c)** spatio-temporal segmentation obtained after applying the strong rule (5 regions); **(c)** final spatio-temporal segmentation (3 regions).