MDL-BASED COHERENT MOTION SEGMENTATION FOR SEMANTIC OBJECT IDENTIFICATION

R. Leonardi, P. Migliorati, G. Tofanicchio

DEA University of Brescia, Brescia, ITALY, E-mail: leon@ing.unibs.it

ABSTRACT

The extraction of semantically meaningful objects that can describe a video sequence provides great expectations for video compression and retrieval. In this paper we propose an effective region merging technique that coherently merge moving image regions in a Minimum Description Length framework. This technique uses a reliability measure that indicates to what extent an affine parameter set represent the motion of an image region. To overcome the region-based motion estimation and segmentation chicken and egg problem, the motion field estimation and the segmentation task are treated separately. A preliminary motion field estimation is carried out starting from a traslational motion model. Concurrently, a Markov Random Field model based algorithm provides for an initial static image partition. For each image region of this spatial segmentation, affine motion parameters are then extracted from the motion field by means of a robust linear regression algorithm. A merging stage based on the proposed technique gives the final spatio-temporal segmentation.

1. INTRODUCTION

In the context of current and future developments of interactive multimedia services, the representation of video sequences as collection of moving objects is going to play an important role [1], [2].

For many transmission and storage applications, significant data compression may be achieved by well-accepted MPEG1, MPEG2 and H.26x standards, based on simple block transform techniques. These block-based representation, however, do not encode data in ways that are suitable for new content-based functionalities, such as interactive television, video-phone, mobile communication, as investigated by new video standards as MPEG4 [3].

Object-based representation at a higher structured semantic level provides great expectation for new search and editing functionalities [4]. Typically, in indexing and retrieval application, a video sequence is subdivided in time into a set of shorter segments, each of which contains similar content. A higher level of abstraction can be achieved if the motion and action of the objects is described within the segment.

However, all the standardization procedure deals with already well-defined and stable semantic objects, and not with the spatio-temporal segmentation that provides these, which still remains an unresolved problem in image analysis.

In this application context, our task is to identify image regions of coherent motions, which are the basis for the successive identification of semantic objects. Motion is in fact one of the most important characteristics to identify objects in a scene.

In [5], given the motion information, regions with the same affine motion are assumed as belonging to the same object. The motion parameters are extracted from the optical flow field by means of linear regression, and temporal segmentation is obtained by clustering in the parameter space; however the resulting motion based image segmentation suffers poor accuracy on object boundaries.

In [6], starting from a luminance-based segmentation, local motion estimation is performed. Then a bottom up segmentation procedure merges coherent adjacent regions. The drawback of these methods is that the parameterisation of the motion field need not have a unique solution.

In [7], [8], starting from a luminance-based segmentation, a merging procedure, based on the mean square error value of the Displaced Frame Difference (DFD) and Motion Difference (MD) respectively, provides the final spatio-temporal segmentation. The use of the DFD alone as a characterization of motion coherence between regions may lead to a significant over segmentation, due to the important value of the DFD in textured areas, while the use of MD alone cannot face a potential initial motion estimator failure in an estimation which uses an indirect parametric estimator.

In [9], the initial set of regions are compared on the basis of a similarity measure that integrates the motion information available in the affine parameter space and the displaced frame difference information.

In [10], a Minimum Description Length (MDL) based formulation for moving object segmentation is proposed; nevertheless, the use of the Displaced Frame Difference (DFD) alone as a characterization of motion coherence between regions may lead to a significant oversegmentation, due to the important value of the DFD in the textured areas.

To improve the performance of these merging criteria, in [11] we introduced a motion field segmentation technique followed by a motion-and-intensity based regularization step which refines the poorly accurate object boundaries. To further improve the segmentation quality towards the identification of semantic object, in this paper we propose a new scheme to robustly segment generic video sequences in a *Minimum Description Length* (MDL) principle [10]. The use of the MDL principle allows to determine the regions motion coherence level by minimizing the number of bits it takes to encode the observed data and the motion model, together with its parameters, which describe the object motion and shape. This feature is intuitively appealing when the goal of motion segmentation is source coding. The merging criteria is based on a reliability measure takes into account both the DFD and the Motion Difference (MD) as criterion of motion coherence.

The paper is organized as follows. Section 2 describes the motion estimation and static segmentation algorithms. The proposed region merging technique is introduced in section 3. Results and conclusions are discussed in the final sections.

2. MOTION ESTIMATION AND STATIC SEGMENTATION

In natural scenes, where changes in camera position, orientation and focal length continuously occur, a global motion compensation is very important in the estimation of "physical" motion fields. Global motion parameters are therefore evaluated.

After global motion compensation, the motion field is estimated by means of a block matching technique. The algorithm exploits the spatial and temporal coherence characteristics of physical motion fields and provides a very smooth estimated motion field with a reduced computational complexity [11].

The initial image partition results from a generalized *k-means clustering* algorithm where spatial constraints are included by the use of a MRF model [12].

3. THE MERGING TECHNIQUE

The proposed algorithm introduces a reliability measure that indicates to what extent an affine motion parameter represents the motion of one image region; it takes into account both the displaced frame difference (DFD) and the motion difference (MD), i.e., the motion error between the estimated motion vector and the motion vector generated by the parametric motion model of the region. The reliability measure associated to a motion parameter θ_j when applied to an image region R_i is

$$\mathbf{C}(i,j) = (\frac{N_i}{2}\log_2(\frac{DFD(i,j)}{N_i}), N_i \log_2(\frac{MD(i,j)}{2N_i})) \quad (1)$$

$$DFD(i,j) = \sum_{\mathbf{x}\in R_i} [I_t(\mathbf{x}) - I_{t-1}(\mathbf{x} - \mathbf{d}_{\theta_j}(\mathbf{x}))]^2 \quad (2)$$

$$MD(i,j) = \sum_{\mathbf{x}\in R_i} ||\mathbf{d}(\mathbf{x}) - \mathbf{d}_{\theta_j}(\mathbf{x})||^2 \quad (3)$$

where N_i is the pixel number of region R_i, I_t is the luminance of the frame at time t,
$\mathbf{d}(\mathbf{x}) = (d_x(\mathbf{x}), d_y(\mathbf{x}))$ is the estimated motion vector at pixel $\mathbf{x} = (x,y)$ and $\mathbf{d}_{\theta_j}(\mathbf{x}) = (d_{x,\theta_j}(\mathbf{x}), d_{y,\theta_j}(\mathbf{x}))$ is the motion vector generated at pixel $\mathbf{x}$ by the affine motion parameter vector θ_j.

In the view of the MDL principle, this reliability measure is proportional to the number of digits it takes to write down the luminance value of the actual frame given the previous frame (DFD) and the motion field (MD) inside region R_i given the parameter vector of the affine motion model. This reliability measure can be exploited as a criterion for a motion based region merging. Given an initial set of image regions with associated motion parameters, the reliability measure is used to construct an adjacency directed graph when the nodes represent the regions and directed arcs are weighted with the set of description length reduction when the motion parameter vector associated to the node the arc starts from is applied to the region the arc is directed towards. Given two adjacent regions R_i, R_j with associated affine motion parameters θ_i, θ_j, as shown in Fig. 1, in the clustering strategy two hypothesis H_1 and H_2 are tested. H_1 means that R_i, R_j are to be merged whereas H_2 means that R_i, R_j are separated. The MDL-test is based on the affine parameter reliability measure as follows. Given the two arcs connecting each pair of nodes, a term by term sum of the description length values of (DFD, MD) is carried out

$$\begin{aligned} \mathbf{C}_H = [L_{ij}, D_{ij}] &= \{\mathbf{C}(i,j) - \mathbf{C}(i,i)\} + && (4) \\ &+ \{\mathbf{C}(j,i) - \mathbf{C}(j,j)\} && (5) \end{aligned}$$

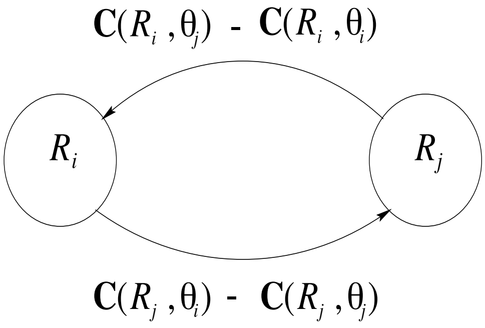
$\mathbf{C}(R_i,\theta_j)$ - $\mathbf{C}(R_i,\theta_i)$
R_i
R_j
$\mathbf{C}(R_j,\theta_i)$ - $\mathbf{C}(R_j,\theta_j)$

Figure 1: Region merging criterion.

This indicator is chosen as the coding cost reduction which determines the level of motion coherence between the two regions; if it is lower than a threshold T_m ispired by the MDL principle, the regions are merged:

$$\min_{\{L_{ij},D_{ij}\}} \mathbf{C}_H = \begin{cases} \leq T_m & H_1 \text{ is verified} \\ > T_m & H_2 \text{ is verified} \end{cases} \tag{6}$$

with

$$T_m = \frac{1}{2} h \log_2(N_p) \tag{7}$$

where h is proportional to the model complexity (i.e., number of regions) and N_p is the number of pixels inside the image. The graph-based region clustering [9], [10] is carried out by applying the decision rule. The two connected nodes which show the greatest motion coherence reliability are merged. The graph weights are then updated. This procedure is iterated until all the description lengths are lower than zero.

4. SIMULATION RESULTS

The proposed region merging technique has been tested on the CIF sequences “Flower Garden”, “Foreman” and “Table Tennis”. As mentioned before, the proposed algorithm expects an initial set of regions as input. The initial static segmentation for the sequence “Foreman” is shown in Fig. 2, while in Fig. 3 the final segmentation is shown where the two motion coherent regions identify in a meaningful way the man and the building in the background. Similarly, starting from the initial segmentation of a “Flower Garden” frame, shown in Fig. 4, the proposed method merges these regions to form meaningful objects. Fig. 5 shows the final segmentation. Figs. 6, 7, show the results for the sequence “Table Tennis”, where the arm and the ball can be used to represent the image.

5. CONCLUSIONS

In this work the problem of spatio-temporal segmentation of video sequences is addressed to identify moving objects in a scene. An affine motion model reliability measure is proposed as a criterion to coherently merge moving image regions. The simulation results show the effectiveness of the proposed region merging technique.

6. REFERENCES

[1] H. Zhang, J. Y. A. Wang, Y. Altunbasak, “Content-Based Video Retrieval and Compression: A unified solution”, in Proc. ICIP-97, pp. 13-16, Santa Barbara, California, Oct. 1997.

[2] D. Zhong, S.-F. Chang, “Spatio-temporal Video Search Using the Object Based Video Representation”, in Proc. ICIP-97, pp. 21-24, Santa Barbara, California, Oct. 1997.

[3] E. Francois, J.-F. Vial and B. Chapeau, “Coding Algorithm with Region-Based Motion Compensation”, *IEEE Transactions on Circuits and System for Video Technology*, Vol. 7, No. 1, pp. 97-108, 1997.

[4] MPEG Requirements Group. MPEG-7: Requirements. ISO/IEC JTC1/SC29/WG11 N2641, MPEG98, Atlantic City, USA, October 1998.

[5] J. Y. Wang, E. H. Adelson, “Representing Moving Images with Layers”, *IEEE Transactions on Image Processing*, Vol. 3, No. 5, pp. 625-638, Sept. 1994.

[6] R. Lancini, M. Ripamonti, S. Tubaro, P. Vicari “Accurate Motion Interpolation by Using a Region Based Motion Estimator”, in Proc. EUSIPCO’98, pp. 1549-1552, Rhodes, Greece, Sept. 1998.

[7] J. G. Choi, S.-W. Lee, S.-D. Kim, ”Spatio-Temporal Video Segmentation Using a Joint Similarity Measure”, *IEEE Trans. on Circuits and System for Video Technology*, Vol. 7, No. 2, pp. 279-286, Apr. 1997.

[8] F. Morier, J. Benois-Pineau, D. Barba, H. Sanson, ”Robust Segmentation of Moving Image Sequences”, in Proc. IEEE ICIP-97, pp. 719-722, Santa Barbara, California, Oct. 1997.

[9] F. Moscheni, S. Bhattacharjee, ”Robust region Merging for Spatio-Temporal Segmentation”, in Proc. IEEE ICIP-96, pp. 501-504, Lausanne, Switzerland, Sept. 1996.

[10] H. Zheng, S. D. Blostein, “Motion-Based Object Segmentation and Estimation Using the MDL Principle”, *IEEE Transactions on Image Processing,* Vol. 4, No. 9, pp. 1223-1235, Sept. 1995.

[11] R. Leonardi, P. Migliorati, G. Tofanicchio, ”A Cooperative Top-Down/Bottom-Up Technique for Motion Field Segmentation”, in Proc. EUSIPCO’98, pp. 1553-1556, Rhodes, Greece, Sept. 1998.

[12] T. N. Pappas, “An Adaptive Clustering Algorithm for Image Segmentation”, *IEEE Transactions on Signal Processing,* Vol. 40, No. 4, pp. 901-914, Apr. 1992.

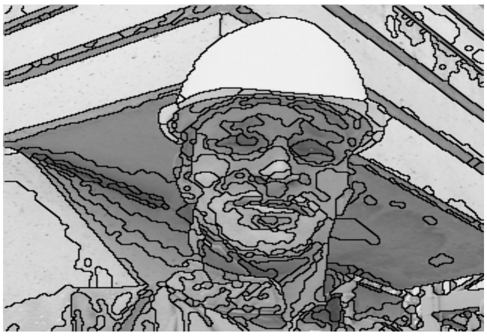

Figure 2: “Foreman” static segmentation.

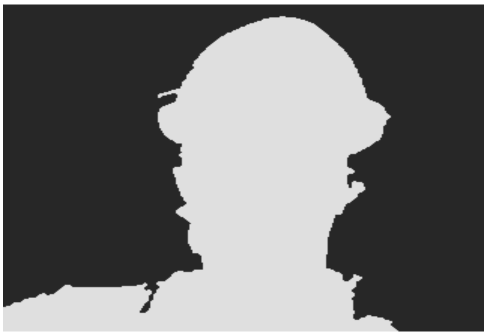

Figure 3: “Foreman” final segmentation.

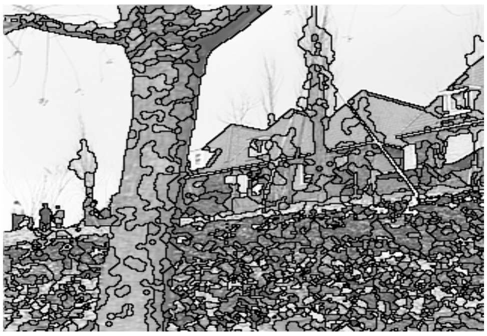

Figure 4: “Flower Garden” static segmentation.

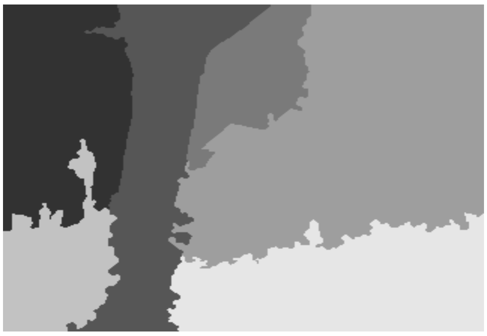

Figure 5: “Flower Garden” final segmentation.

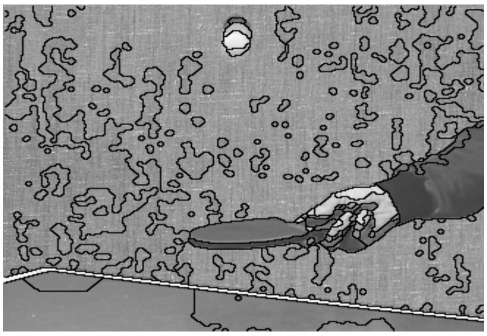

Figure 6: “Table tennis” static segmentation.

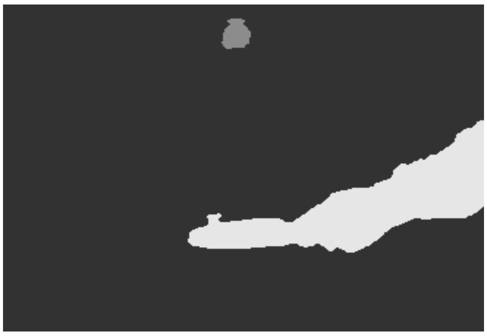

Figure 7: “Table tennis” final segmentation.