A COOPERATIVE TOP-DOWN/BOTTOM-UP TECHNIQUE FOR MOTION FIELD SEGMENTATION

R. Leonardi, P. Migliorati, G. Tofanicchio
University of Brescia, DEA, via Branze, 38, 25123, Brescia, Italy
Tel. +39-30-3715433, Fax. +39-30-380014
e-mail: pier@ing.unibs.it

ABSTRACT

The segmentation of video sequences into regions underlying a coherent motion is one of the most useful processing for video analysis and coding. In this paper, we propose an algorithm that exploits the advantages of both top-down and bottom-up techniques for motion field segmentation. To remove camera motion, a global motion estimation and compensation is first performed. Local motion estimation is then carried out relying on a traslational motion model. Starting from this motion field, a two-stage analysis based on affine models takes place. In the first stage, using a top-down segmentation technique, macro-regions with coherent affine motion are extracted. In the second stage, the segmentation of each macro-region is refined using a bottom-up approach based on a motion vector clustering. In order to further improve the accuracy of the spatio-temporal segmentation, a Markov Random Field (MRF)-inspired motion-and-intensity based refinement step is performed to adjust objects boundaries.

1 INTRODUCTION

Motion is one of the most important characteristics to identify objects in a scene. Motion based segmentation is therefore very important in applications such as dynamic scene analysis, time-to-collision calculation, obstacle detection and tracking of moving objects. Moreover, a motion-based segmentation can be directly used also in hybrid video coding architectures [1]. Contrarily to an intensity-based approach [2], motion based segmentation deals with few and large regions that are likely to identify real moving objects in a scene.

As motion estimation and segmentation are interdependent, they should be carried out jointly. The algorithms proposed in literature to solve this ill-posed problem can be divided in two classes: bottom-up and top-down algorithms.

In [3], given the motion information, regions with the same affine model parameters are assumed as belonging to the same object. These parameters are extracted from the optical flow field by means of linear regression, and temporal segmentation is obtained by clustering in the affine parameter space. The resulting motion based image segmentation suffers poor accuracy on object boundaries. In [4] a top-down hierarchical motion segmentation and estimation scheme is proposed. First, the dominant motion is estimated; the current image is then compared with the motion compensated one, and new regions are defined as the areas corresponding to large prediction errors. As a new region has been detected, its boundary is refined by superimposing the results of luminance based segmentation. The same procedure is recursively applied to every new detected object.

In this paper, we propose a cooperative approach for motion field segmentation that gets rid of the drawbacks of the bottom-up techniques by means of the advantages of the top-down techniques, and vice-versa. A global/local motion estimation and compensation is first performed relying on a traslational motion model. Starting from this motion field, macro-regions with coherent affine motion are extracted by means of a top-down technique. As the dominant motion assumption may rise to an under-segmented image, the motion segmentation of each macro-region is refined using a bottom up approach. This second stage works on a limited, highly motion coherent set of displacement vectors, so that only few clusters are needed and the possible misunderstanding of pixel classification is limited. The moving objects boundaries are finally adjusted minimizing a Markov Random Field (MRF)-inspired motion-and-intensity based function.

The paper is organized as follow. Section 2 describes the adopted global/local motion estimation algorithm. Section 3 is devoted to the description of the proposed motion segmentation algorithm whereas the segmentation refinement algorithm is presented in Section 4. Simulation results and conclusions are given in the final Sections.

2 GLOBAL/LOCAL MOTION ESTIMATION

In natural scenes, where changes in camera position, orientation and focal length may occur continuously, a global motion compensation is very important in the estimation of “physical” motion fields. Global motion parameters are therefore evaluated as described in [5]. After global motion compensation, the local motion field is estimated by means of a block matching technique [6]. The algorithm exploits the spatial and temporal coherence characteristics of physical motion fields and provides a very smooth motion field with a reduced computational complexity.

3 COOPERATIVE MOTION FIELD SEGMENTATION

The proposed algorithm is based on two stages: a top-down and a bottom-up motion field segmentation. The top-down stage, which does not need any initial segmentation to start with nor requires any assumption on the number of regions in the scene, provides for a first motion field segmentation. The bottom-up approach, which does not assume the presence of any dominant motion, refines the regions characteristics.

3.1 Top-Down Motion Field Segmentation

In the top-down approach, video scenes are hierarchically segmented into several differently moving objects. Unlike the approach presented in [4], where segmentation and motion estimation are treated as combined, in our approach, moving objects are extracted from a given estimated motion field.

Let $d(\mathbf{x}) = (d_x(\mathbf{x}), d_y(\mathbf{x}))$ be the estimated motion vector at pixel $\mathbf{x} = (x, y)$ and $d_\theta(\mathbf{x}) = (d_{x,\theta}(\mathbf{x}), d_{y,\theta}(\mathbf{x}))$ be the motion vector generated at pixel $\mathbf{x}$ by the affine motion parameter vector $\theta = (a_{x0}, a_{xx}, a_{xy}, a_{y0}, a_{yx}, a_{yy})$ where

$$\begin{aligned} d_{x,\theta}(\mathbf{x}) &= a_{x0} + a_{xx}x + a_{xy}y \\ d_{y,\theta}(\mathbf{x}) &= a_{y0} + a_{yx}y + a_{yy}y. \end{aligned}$$

First, stationary regions are detected and removed by means of a thresholding process on the motion vectors; the remaining connected moving regions form the set $\mathcal{R}$ of first hierarchical level objects. To identify the dominant motion from a given set of motion vectors, a “Weighted Least Squares” method is adopted [7]. For every object $R \in \mathcal{R}$, an affine motion parameter vector θ_R^0 is initially estimated by means of a least square procedure. The residual error between the actual motion vector $d(\mathbf{x})$ and the displacement $d_{\theta_R^0}(\mathbf{x})$ given by the estimated motion parameters θ_R^0 is calculated for every pixel $\mathbf{x}$ in the region R as

$$\begin{aligned} \epsilon_{x,\theta_R^0}(\mathbf{x}) &= d_x(\mathbf{x}) - d_{x,\theta_R^0}(\mathbf{x}), \\ \epsilon_{y,\theta_R^0}(\mathbf{x}) &= d_y(\mathbf{x}) - d_{y,\theta_R^0}(\mathbf{x}). \end{aligned}$$

Hence, a *robust* estimation of these residual errors standard deviation σ_R is carried out [7]. The pixels whose motion vector shows a residual error greater than $h\sigma_R$ (h ranges from 3 to 1) are considered as *outliers* and they are assigned to the next hierarchical level. Each *inlier* pixel is then allocated a weight which is inversely proportional to the residual error

$$w_k(\mathbf{x}) = [1 - (E_{\theta_R^{k-1}}(\mathbf{x}))/(h\sigma_R)^2]^2$$

where

$$E_{\theta_R^k}(\mathbf{x}) = \|d(\mathbf{x}) - d_{\theta_R^k}(\mathbf{x})\|^2. \quad (1)$$

At the k^{th} iteration the new set of affine parameters for the *inlier* pixels of object R is obtained by weighted least squares. The process ”weighting coefficients calculation - weighted least squares estimation” is then iterated until the number of motion vectors inside the region reaches an asymptotic value. This procedure is performed for every hierarchical level (formed by the previous hierarchical level *outlier* pixels), until all pixels are classified.

At the end, this motion segmentation step provides a partition $\mathcal{L}$ of the image into N_L macro-regions, where $N_L = \sum_{R\in\mathcal{R}} L_R$ and L_R is the hierarchical level number of each object $R \in \mathcal{R}$, as such hierarchical number represents the number of regions that when combined form object R. The top-down approach often provides an under-segmented image, due to the dominant motion assumption. This draft segmentation $\mathcal{L}$ is the initial guess for the bottom-up step.

3.2 Bottom-Up Motion Field Segmentation

A motion vector k-means clustering algorithm based on an MRF model [8] is used in this step.

For every macro-region $L \in \mathcal{L}$, pixels with similar motion vectors are clustered. Motion vectors whose associated regions have a size greater than a suitable threshold T are then selected as cluster centres. In this way the problem of choosing the cluster number K_L, $L \in \mathcal{L}$, is automatically solved. Moreover, for each macro-region, only few clusters are considered. If only one cluster is selected, no region splitting is needed; the macro region corresponds to a coherently moving object and the associated motion parameter is reliable. If more than one cluster is identified, a modified k-means clustering method is carried out on the macro-region motion vectors. An MRF model is exploited to include spatial connectivity among elements in the same class. The region labelling is obtained by minimizing the following energy function:

$$\min_{k=1,\ldots,K_L;\mathbf{x}\in L} U_1(\mathbf{x}) = E_{\theta_k}(\mathbf{x}) + \sum_{\mathbf{y}\in\eta_1(\mathbf{x})} V_1(\mathbf{x},\mathbf{y}) \quad (2)$$

$$V_1(\mathbf{x},\mathbf{y}) = \begin{cases} -\beta_1 & \text{if } l_\mathbf{x} = l_\mathbf{y} \\ +\beta_1 & \text{if } l_\mathbf{x} \neq l_\mathbf{y} \end{cases} \quad (3)$$

where $\eta_1(\mathbf{x})$ is the first-order neighbourhood (i.e., 4-connected set of points) of pixel $\mathbf{x}$ and $l_\mathbf{x}, l_\mathbf{y}$ are labels at pixels $\mathbf{x}$ and $\mathbf{y}$. The first term of Eq. 2 measures the fitting of the motion parameter vector θ_k of region k to the observed motion vector $d(\mathbf{x})$, while the second one accounts for the region spatial connectivity. The Iterated Conditional Modes (ICM) algorithm provides the maximum *a posteriori* (MAP) estimation of the motion segmentation [9].

At the image level, i.e., for all formed regions, a clustering in the parameter space [3] is then carried out in order to merge coherently moving regions which were possibly separated after the two motion segmentation steps as they were assigned to different hierarchical levels. Small residual regions are finally merged into the most coherently moving surrounding object, i.e., to the region which provides the minimum sum of residual errors (Eq. 1). This procedure provides a set $\mathcal{M}$ of N_M meaningful regions characterized by an accurate set of affine motion parameters.

4 SPATIO-TEMPORAL SEGMENTATION REFINEMENT

The final region boundaries identification is achieved through an MRF-based regularization approach [8], [10]. A weighted sum of displaced frame difference (which accounts for motion) and intensity difference (which accounts for luminance values of the region) [11], is proposed as joint similarity measure between boundary pixels and regions. The motion of each region M is identified by a set of affine parameters θ_M, which have been estimated in the motion field segmentation steps. The motion similarity between the pixel $\mathbf{x}$ under consideration and the region $M \in \mathcal{M}$ is defined as

$$S_{m,\theta_M}(\mathbf{x}) = |I_t(\mathbf{x}) - I_{t-1}(x - d_{x,\theta_M}(\mathbf{x}), y - d_{y,\theta_M}(\mathbf{x}))|^2$$

where $I_t(\mathbf{x})$ is the grey value at $\mathbf{x}$ at time t. The intensity of each region M is identified by a set of three parameters $\phi_M = (\alpha, \beta, \gamma)$, which are estimated by means of a linear regression method on luminance values. Hence the intensity similarity is defined as

$$S_{i,\phi_M}(\mathbf{x}) = |I(\mathbf{x}) - I_{\phi_M}(\mathbf{x})|^2$$

where $I_{\phi_M}(\mathbf{x}) = \alpha + \beta x + \gamma y$ is the grey value generated at pixel $\mathbf{x}$ by the luminance parametes ϕ_M of region M. To account for local intensity variations in the image, the luminace parameters are estimated over samples that lie into a $d \times d$ window centered around the current pixel $\mathbf{x}$.

Thus, a new joint similarity measure can be defined as the weigthed sum of motion similarity plus the intensity similarity

$$S_M(\mathbf{x}) = \mu S_{m,\theta_M}(\mathbf{x}) + (1 - \mu) S_{i,\phi_M}(\mathbf{x})$$

where μ is a weight factor.

The adjustment is carried out locally at the boundary of each region assigning the processed pixel to the region M whose motion and luminance parameters provides the best similarity value. The energy function to be minimized is

$$\min_{M \in \mathcal{M}'} U_2(\mathbf{x}) = S_M(\mathbf{x}) + \sum_{\mathbf{y} \in \eta_2(\mathbf{x})} V_2(\mathbf{x}, \mathbf{y})$$

where $\mathcal{M}'$ is the set of candidate labels, i.e., labels at $\mathbf{x}$ and its second-order neighbourhood $\eta_2(\mathbf{x})$ (i.e., 8-connected set of points), and $V_2(\mathbf{x}; \mathbf{y})$ represents the MRF potential function whose expression is given by Eq. 3 with index 2 replacing index 1. The ICM algorithm [9] is used to obtain the MAP estimate of the final spatio-temporal segmentation $\mathcal{S}$.

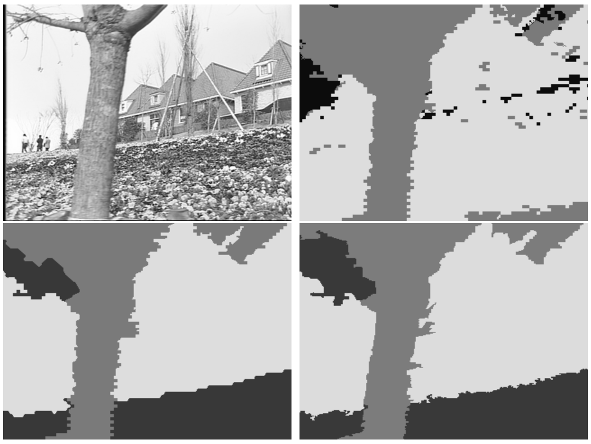

Figure 1: "Flower Garden": (top left) one frame; (top right) Top-Down segmentation $\mathcal{L}$ ($N_L = 2$ + stationary background); (bottom left) Bottom-Up segmentation $\mathcal{M}$ ($N_M = 3$) with $T = 200$, $\beta_1 = 0.5$; (bottom right) Spatio-Temporal segmentation $\mathcal{S}$ with $\beta_2 = 128$, $d = 20$, $\mu = 0.5$.

5 SIMULATION RESULTS

The proposed motion field segmentation technique has been tested on the CIF sequences "Flower Garden" (Fig.1), "Table Tennis" (Fig.2) and "Foreman" (Fig.3). Fig. 1, top right, shows "Flower Garden" top-down image partition $\mathcal{L}$ into $N_L = 2$ macro-regions L (the tree and the flower-bed/houses) and the stationary background (black). The flower-bed and the houses are detected in the bottom-up stage (Fig.1, bottom left), where the motion based partition $\mathcal{M}$ gives rise to $N_M = 3$ coherently moving regions M. Fig.1, bottom right, shows the spatio-temporal segmentation where the boundary "blockiness" effect has been removed by the regularization step. "Table Tennis" sequence (Fig.2, bottom right) is effectively segmented ($N_M = 3$ objects, i.e., the ball, the arm, the racket with the hand, and the background) with precisely located motion boundaries. Even if "Foreman" sequence (Fig.3, bottom right) results in an over-segmentation, the final object detection is still meaningfull.

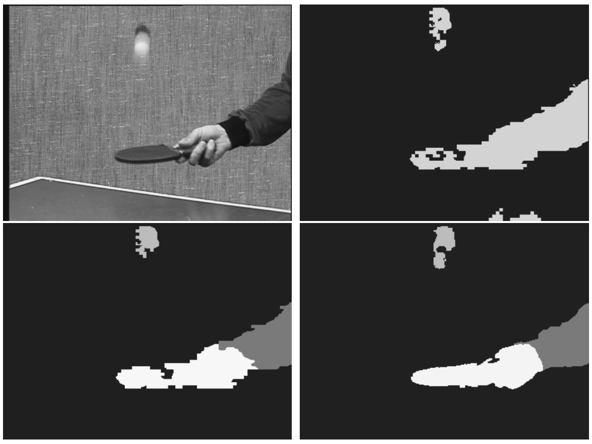

Figure 2: "Table Tennis": (top left) one frame; (top right) Top-Down segmentation $\mathcal{L}$ ($N_L = 3$ + stationary background); (bottom left) Bottom-Up segmentation $\mathcal{M}$ ($N_M = 3$) with $T = 200$, $\beta_1 = 0.5$; (bottom right) Spatio-Temporal segmentation $\mathcal{S}$ with $\beta_2 = 64$, $d = 20$, $\mu = 0.5$.

6 CONCLUSIONS

The segmentation of video sequences into coherent moving objects has been addressed in this paper. The advantages of the top-down and bottom-up motion field segmentation algorithms have been exploited to propose a new cooperative method. Furthermore, an intensity-and-motion based regularization step has been suggested to get an accurate spatio-temporal segmentation. Simulation results show that moving objects are effectively detected and their boundaries are accurately located.

Future developments will be devoted to implement a more robust cluster number selection and to introduce a tracking algorithm to maintain segmentation temporal stability.

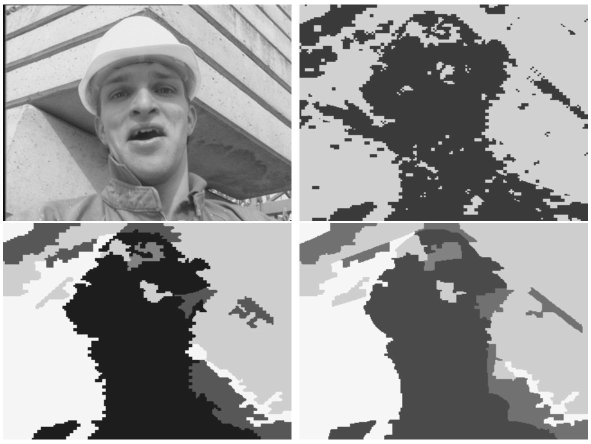

Figure 3: "Foreman": (top left) one frame; (top right) Top-Down segmentation $\mathcal{L}$ ($N_L = 2$); (bottom left) Bottom-Up segmentation $\mathcal{M}$ ($N_M = 4$) with $T = 200$, $\beta_1 = 0.5$; (bottom right) Spatio-Temporal segmentation $\mathcal{S}$ with $\beta_2 = 256$, $d = 20$, $\mu = 0.5$.

References

[1] R. Leonardi, "Region Based Image and Video Compression", in *Proc. 1995 Int. Conf. on Digital Signal Processing*, Cyprus, Jun. 1995.

[2] P. Salembier, M. Pardas, "Hierarchical Morphological Segmentation for Image Sequence Coding", *IEEE Trans. on Image Processing*, Vol. 3, No. 5, pp. 639-651, Sept. 1994.

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

[4] N. Diehl, "Object-Oriented Motion Estimation and Segmentation in Image Sequences", *Signal Processing: Image Communication*, Vol. 3, No. 1, pp. 23-56, 1991.

[5] P. Migliorati, S. Tubaro, "Multistage Motion Estimation for Image Interpolation", *Signal Processing: Image Communication*, Vol. 7, No. 3, pp. 187-199, 1995.

[6] G. de Haan, H. Huijgen, "New Algorithm for Motion Estimation", in *Proc. Third Int. Workshop on HDTV*, Torino, Italy, 1989.

[7] P. Meer, D. Mintz, A. Rosenfeld, "Robust Regression Methods for Computer Vision: A Review". *International Journal of Computer Vision*, Vol. 6, No. 1, pp. 59-70, 1991.

[8] F. Pedersini, A. Sarti, S. Tubaro, "Combined Motion and Edge Analysis for a Layered Based Representation of Image Sequences", in *Proc. IEEE-ICIP '96*, Lausanne, Switzerland, pp. 921-924, Sept. 1996.

[9] J. Besag, "On the Statistical Analysis of Dirty Pictures", *Journal of Royal Statistical Society* B 48, No. 3, pp. 259-302, 1986.

[10] J. Konrad, V.-N. Dang, "Coding-Oriented Video Segmentation Inspired by MRF Models", in *Proc. IEEE-ICIP '96*, Lausanne, Switzerland, pp. 909-912, Sept. 1996.

[11] 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.