SPATIO-TEMPORAL SEGMENTATION AND REGIONS TRACKING OF HIGH DEFINITION VIDEO SEQUENCES BASED ON A MARKOV RANDOM FIELD MODEL

Olivier Brouard, Fabrice Delannay, Vincent Ricordel and Dominique Barba

University of Nantes – IRCCyN – UMR CNRS 6597 – IVC team
Polytech' Nantes, rue Christian Pauc, BP 50609, 44306 Nantes, France
olivier.brouard@univ-nantes.fr

ABSTRACT

In this paper[1], we propose a Markov Random Field sequence segmentation and regions tracking model, which aims at combining color, texture, and motion features. First a motion-based segmentation is realized. Namely the global motion of the video sequence is estimated and compensated. From the remaining motion information, a rough motion segmentation is achieved. Then, we use a Markovian approach to update and track over time the video objects. The spatio-temporal map is updated and compensated using our Markov Random Field segmentation model to keep consistency in video objects tracking.

Index Terms— Video Motion-Based Segmentation, Markov Random Fields, Regions Tracking.

1. INTRODUCTION

Image segmentation and video objects tracking are the subjects of large researches for video coding. For instance, the new video standard H.264 allows a wide choice of coding strategies, one possible is to use adapted coding parameters for the video object during several frames.

To track spatio-temporal objects in a video sequence, they need to be segmented. By video object, we mean typically, a spatio-temporal shape characterized by its texture, its color, and its own motion that differs from the global motion of the shot. In the literature, several kinds of methods are described, they use spatial and/or temporal [1] information to segment the objects. In the case of spatial information, good segmentation results have been obtained using Markov Random Fields (MRF) [2, 3]. Indeed, the MRF define a class of statistical models which enable to describe both the local and global properties of segmentation maps. The methods based on temporal information need to know the global motion of the video to perform an effective video objects segmentation. Horn and Schunck [4] proposed to determine the optical flow between two successive frames. Otherwise, the motion parametric model of the successive frames can be estimated [5]. Once the motion model is known, the global motion is back-compensated, and only the moving objects remain with their local motion information. Studies in motion analysis have shown that motion-based segmentation would benefit from including not only motion, but also the intensity cue, in particular to retrieve accurately the regions boundaries. Hence the knowledge of the spatial partition can improve the reliability of the motion-based segmentation. As a consequence, we propose a MRF model combining the motion information and the spatial features of the sequence to achieve an accurate segmentation and video objects tracking.

In previous works, we used a motion information per block, and for a group of frames (GOF), to estimate the global motion and achieve the motion-based segmentation [6]. First the method, considering several successive reference frames, estimates the motion of spatio-temporal tubes, with the assumption of an uniform motion along the GOF. Next a motion vectors accumulation permits to estimate robustly the parameters of an affine motion model (the global motion). Finally the global motion is compensated, and the motion segmentation is achieved from the compensated motion vectors. In this paper, we propose a MRF model which aims at combining color, texture, and motion features. This model permits to improve an initial motion-based segmentation, and to compute video objects with accurate boundaries. Moreover the spatio-temporal map from the previous GOF is updated and compensated using our MRF model to proceed and keep consistency in video objects tracking.

In the following section, we briefly present our motion-based segmentation method based on spatio-temporal tubes. In section 3, we describe the MRF sequence segmentation and regions tracking algorithm. Finally, we show the simulation results in section 4, and we conclude in section 5.

2. MOTION-BASED SEGMENTATION

2.1. Motion estimation based on tubes

To extract motion information correlated with the motions of real life objects in the video shot, we consider several successive frames and we make the assumption of an uniform

[1]This research was carried out within the framework of the ArchiPEG project financed by the ANR (convention N°ANR05RIAM01401).

motion between them. Taking account of perceptual considerations, and of the frame rate of the next HDTV generation in progressive mode, we use a GOF composed of 9 frames [6, 7]. The goal is to ensure the coherence of the motion along a perceptually significant duration.

Figure 1 illustrates how a spatio-temporal tube is estimated considering a block of the frame F_t at the GOF center: an uniform motion is assumed and the tube passes through the 9 successive frames such as it minimizes the error between the current block and those aligned.

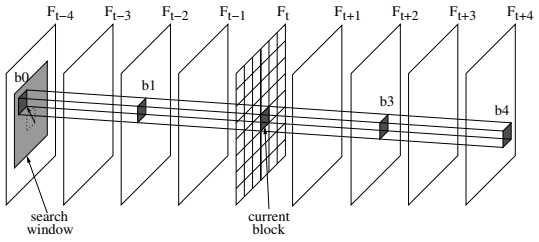

Fig. 1. Spatio-temporal tube used to determine the motion vector of a given block.

We get a motion vectors field with one vector per tube, and one tube for each block of the image F_t. This motion vectors field is more homogeneous (smoother) and more correlated with the motion of real life objects, this field is the input of the next process: the global motion estimation.

2.2. Robust global motion estimation

The next step is to identify the parameters of the global motion of the GOF from this motion vectors field. We use an affine model with six parameters. First, we compute the derivatives of each motion vector and accumulate them in an histogram (one respective histogram for each global parameter). The localization of the main peak in the histogram produces the value retained for the parameter. Then, once the deformation parameters have been identified, they are used to compensate the original motion vectors field. Thus, the remaining vectors correspond only to the translation motions. These remaining motion vectors are then accumulated in a two dimensions (2D) histogram. The main peak in this 2D histogram represents the values of the translation parameters (for more details, the readers are invited to see our previous work [6]).

2.3. Motion segmentation

In the previous 2D accumulation histogram used to estimate the global motion translation, we assume that each peak represents an object motion, so we do not retain only the main peak but all of them to segment the GOF.

For all the positions connected to the main peak, a local gradient is computed. All the connected cells, for which the gradient is positive, are considered as belonging to the peak. Then, a new maximum is detected among all the remaining (not labeled) cells, and the algorithm is iterated as long as there remain non null cells without label. If one cell is labeled as belonging to several peaks, it is linked to the closest peak. We get here an rough segmentation map per GOF. Our goal becomes, using a MRF, to improve those initial segmentation maps and to link them temporally.

3. MARKOV RANDOM FIELD MODEL

We express the markovian proprieties of a field by an explicit distribution.

Let $E = \{E_s,\ s \in S\}$ be the label field defined on the lattice S of sites s, in our case each site is associated with a tube, and the sites of a segmented region (corresponding to a moving object through successive GOF) are labelled similarly. Let $O = \{O_s,\ s \in S\}$ be the observation field. Realizations of fields E (respectively O) will be denoted $e = \{e_s,\ s \in S\}$ (respectively $o = \{o_s,\ s \in S\}$). Let Λ (respectively Ω) be the set of all possible realizations of E (respectively label configurations e). With respect to the chosen neighborhood system $\eta = \{\eta_s,\ s \in S\}$, $(E,\ O)$ is modeled as a MRF. The optimal label field $\hat{e}$ is derived according to the *Maximum A Posteriori* (*MAP*) criterion. The Hammersley-Clifford theorem [8] established the equivalence between Gibbs distribution and the MRF, the optimal label configuration is then obtained by minimizing a global energy function $U(o,\ e)$:

$$\hat{e} = \arg\min_{e \in \Omega} U(o,\ e) \tag{1}$$

Due to the Markovian property of the field, the energy function is written as the sum of elementary potential functions defined on locally structures called *cliques* [9]:

$$U(o,\ e) = \sum_{c \in C} V_c(o,\ e), \tag{2}$$

where C is the set of *cliques* from S associated to the neighborhood η. The potential function V_c is locally defined on the *clique* c and gives the local interactions between its different elements. The form of the potential function V_c is problem dependent and defines its local and global properties.

3.1. Potential functions

Considering one GOF, a segmented region should respect a spatial coherence, it means that the segmented region (constituted of tubes) should be locally homogeneous and compact. The corresponding potential function is related to a Markov model associated to an eight-neighborhood system. The model favours spatially homogeneous regions, by the choice of the potential function:

$$\forall t \in \eta_s \begin{cases} V_{c_s} = \beta_s & \text{if } e_t \neq e_s, \\ V_{c_s} = -\beta_s & \text{if } e_t = e_s, \end{cases}$$

with $\beta_s > 0$. In our case, C is the the set of spatial second order *cliques*. Each *clique* corresponds to a pair of neighbor-

ing and connected tubes:

$$W_1(e) = \sum_{c_s \in C_s} V_{c_s}(e_s,\, e_t),$$

where C_s represents the set of all the spatial *cliques* of S.

3.1.1. Color features

Inside a GOF, we want to compare the color distributions of a site with the other regions. Many methods are adapted to the discrete case (intersection, L_2, χ_2, ...), we have chosen the Bhattacharyya coefficient based on similarities computation.

The discrete densities of the color distributions of the current site s, $\hat{s} = \{\hat{s}_u\}_{u=1..m}$, and of the region $R(e_s)$ constituted by the sites labeled e_s, $\widehat{R(e_s)} = \left\{\widehat{R(e_s)}_u\right\}_{u=1..m}$, are computed from the color histogram with m bins and considering only the frame F_t at the GOF center. The corresponding Bhattacharyya coefficient is then defined by:

$$\rho_c = \rho_c(\widehat{R(e_s)},\, \hat{s}) = \sum_{u=1}^{m} \sqrt{\widehat{R(e_s)}_u .\, \hat{s}_u}.$$

From this coefficient, we deduce a distance between 0 and 1: $d_c = (1 - \rho_c(\widehat{R(e_s)}, \hat{s}))^{1/2}$. The potential W_2 for the color features is defined as follows:

$$W_2(e_s,\, o_s,\, o(R(e_s))) = \sum_{s \in S} \sqrt{1 - \rho_c(\widehat{R(e_s)},\, \hat{s})}.$$

3.1.2. Texture features

Inside a GOF, in order to compare the image textures, the two different spatial gradients ($\overline{\Delta V}$, $\overline{\Delta H}$) are used, each one is computed for each pixel and each regions of the frame F_t at the GOF center. In practice, we use Sobel filters, and the Bhattacharyya coefficient to compute similarities. Namely, the discrete densities of the texture distributions of the current site s, $\hat{s} = \{\hat{s}_u\}_{u=1..n}$, and the region $R(e_s)$ formed by the sites labelled e_s, $\widehat{R(e_s)} = \left\{\widehat{R(e_s)}_u\right\}_{u=1..n}$, are calculated from the texture histogram with n bins. The Bhattacharyya coefficient for the texture distributions is defined by:

$$\rho_t = \rho_t(\widehat{R(e_s)},\, \hat{s}) = \sum_{u=1}^{n} \sqrt{\widehat{R(e_s)}_u .\, \hat{s}_u}.$$

The potential W_3 for the texture features is given by:

$$W_3(e_s,\, o_s,\, o(R(e_s))) = \sum_{s \in S} \sqrt{1 - \rho_t(\widehat{R(e_s)},\, \hat{s})}.$$

3.1.3. Motion features

Inside a GOF, the main criterion for the segmentation is often the motion: for a given region, the motion vectors of its tubes should have close values. Therefore we want to associate an energy to assess the difference between the motion of a tube and the motion of a region.

In the section 2, we explained how the motion vector of a tube is estimated, and how each region is located thanks to a peak in a 2D accumulation histogram. So the motion vector associated to a peak is also the estimated motion of the region in the GOF. The distance between the motions of a tube, and a region, according to their norms and their directions, follows:

$$d_m = \frac{\overrightarrow{MV_s} \times \overrightarrow{MV_{R(e_s)}}}{max(\left\|\overrightarrow{MV_s}\right\|, \left\|\overrightarrow{MV_{R(e_s)}}\right\|)},$$

where $\overrightarrow{MV_s}$, and $\overrightarrow{MV_{R(e_s)}}$ are respectively the motion vectors of the site s, and of the region $R(e_s)$ formed by the sites labelled e_s . In order to constrain this distance between 0 and 1, we compute $\rho_m(\overrightarrow{MV_{R(e_s)}}, \overrightarrow{MV_s}) = (d_m + 1)/2$. The corresponding potential function W_4 is given by:

$$W_4(e_s, o_s, o(R(e_s))) = \sum_{s \in S} \sqrt{1 - \rho_m(\overrightarrow{MV_{R(e_s)}}, \overrightarrow{MV_s})}.$$

3.1.4. Regions tracking

In order to track the regions between two successive GOF, we compare their segmentation maps. Exactly the segmentation map of the previous GOF, is first compensated using all of the motion information (global motion, motion vectors of its objects). Next we compare the labels of the regions in the previous and in the current GOF. A metric based on the color, the texture, and the recovery between the regions, is used. For the color, and the texture, we adapt the Bhattacharyya coefficients detailed in sub-sections 3.1.1 and 3.1.2. A region of the current GOF takes the label of the closest region of the previous GOF (if their distance is small enough).

The compensated map of the previous GOF is used to improve the current map through the potential function:

$$\begin{cases} V_{c_t} = \beta_t & \text{if } e_s(t) \neq e_s(t-1), \\ V_{c_t} = -\beta_t & \text{if } e_s(t) = e_s(t-1), \end{cases}$$

with $\beta_t > 0$, and where $e_s(t)$, and $e_s(t-1)$ are respectively the labels of the site for the current, and the motion compensated previous GOF. Here C is the set of temporal second order *cliques*. Each *clique* corresponds to a pair of adjacent tubes between the previous and the current GOF:

$$W_5(e(t)) = \sum_{c_t \in C_t} V_{c_t}(e_s(t), e_s(t-1)),$$

where C_t is the set of all the temporal *cliques* of S.

Inside a GOF, when the motion of the potential objects are very similar, the motion-based segmentation failed to detect them. In this case, the initial segmentation map for our MRF segmentation model contains no information, hence, we use the motion compensated map from the previous GOF as initialization for our MRF segmentation model. This process allows to keep consistency for video objects tracking through the sequence GOF.

3.2. Energy minimization

The global energy function $U(o, e)$ is expressed as:

$$U(o, e) = \beta_1.W_1 + \beta_2.W_2 + \beta_3.W_3 + \beta_4.W_4 + \beta_5.W_5,$$

where β_1, β_2, β_3, β_4,and β_5 are respectively the weights for the potential functions W_1, W_2, W_3, W_4, and W_5.

The rough maps obtained from the motion-based segmentation are used as initialization for the optimization process. The tubes located at the borders of the moving objects, or in the uniform areas have the highest probability to be misclassified, they represent the unstable sites. We use a stack of instability to determine the visit order of the unstable sites.

First, we check the stability of each site (i.e. if the energy associated with the current label is minimal). If the energy variation for the site equals zero, $\Delta U(s) = 0$, the site is stable. On the contrary, we compute the energy variation: $\Delta U(s) = U(s, e_c) - U(s, e_s)$, where e_c, and e_s are respectively the current label and the new label of the site s which minimizes the energy. Next a decreasing instability stack is built. Its first site(the most unstable), is updated with the new label which minimizes the energy. The energies of the neighboring sites are modified too, so the instability stack has to be updated at each iteration.

4. SIMULATION RESULTS

We used one 1080p (*Tractor*), and two 720p (*Shields* and *New mobile calendar*) HD sequences from SVT [10]. These video sequences contain one or several moving objects.

Table 1 presents the number of the detected moving objects using only the motion-based segmentation (MBS), and with our MRF model. Although, the method is improved, note that for the *Tractor* sequence it failed to detect all of them. Indeed, at the end of this sequence, the tractor is too small (because of a camera zoom out) to be detected.

Sequence	MBS	MRF model
Tractor	33%	84%
New Mobile Calendar	85%	92%
Shields	94%	100%

Table 1. Ratio of detected moving objects.

Figure 2 shows the segmentation maps using the only MBS (top row), and with our MRF model (bottom row) for three successive GOF of the *Tractor* sequence. The moving objects are correctly detected with the MBS, but the labels between the GOF are incoherent (the same object is labellized differently). With our MRF model, the boundaries of the detected moving objects are more regular than those obtained with the MBS. Moreover, video objects tracking is successful with our MRF model, since the tractor label is the same between the three GOF.

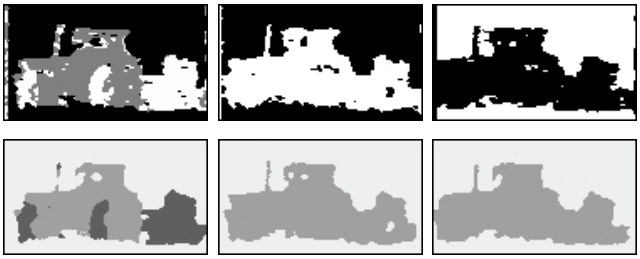

Fig. 2. Segmentation maps and tracking for *Tractor* (GOF 14, 15, 16) using the MBS (top row) and our MRF model (bottom row).

5. CONCLUSIONS

In this paper, we have presented a Markov Random Field (MRF) model to segment and track video objects. Our MRF model combines color, texture, and motion features. First, a motion-based segmentation (MBS) is realized for a GOF of nine frames. Next the MRF model is applied to improve the MBS using spatial features, and to keep consistency between the successive GOF segmentation maps. A video objects tracking is then achieved.

6. REFERENCES

[1] R. Megret and D. DeMenthon, "A survey of spatio-temporal grouping techniques," Tech. Rep., LAMP-TR-094/CS-TR-4403, University of Maryland, 1994.

[2] C. Kervrann and F. Heitz, "A Markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics," *IEEE Trans. on Image Processing*, vol. 4, no. 6, pp. 856 – 862, 1995.

[3] Z. Kato and T.C. Pong, "A markov random field image segmentation model for textured images," *Image and Vision Computing*, vol. 24, pp. 1103 – 1114, October 2006.

[4] B.K.P. Horn and B.G. Schunck, "Determining Optical Flow," *Artificial Intelligence*, vol. 17, no. 1 – 3, pp. 185 – 203, 1981.

[5] J. M. Odobez and P. Bouthemy, "Robust Multiresolution Estimation of Parametric Motion Models," *Journal of Visual Communication and Image Representation*, vol. 6, December 1995.

[6] O. Brouard, F. Delannay, V. Ricordel, and D. Barba, "Robust Motion Segmentation for High Definition Video Sequences using a Fast Multi-Resolution Motion Estimation Based on Spatio-Temporal Tubes," Lisbon, Portugal, in Proc. PCS 2007.

[7] O. Le Meur, P. Le Callet, and D. Barba, "A Coherent Computational Approach to Model Bottom-Up Visual Attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802 – 817, May 2006.

[8] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B*, vol. 36, pp. 192 – 236, 1974.

[9] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721 – 741, 1984.

[10] SVT, "Overall-quality assessment when targeting wide xga flat panel displays," Tech. Rep., SVT corporate development technology, 2002.