SEGMENTATION OF NON-RIGID VIDEO OBJECTS USING LONG TERM TEMPORAL CONSISTENCY

Chaumont Marc, Pateux Stéphane and Nicolas Henri
IRISA, Campus de Beaulieu
35042 Rennes, France
Email: Marc.Chaumont@irisa.fr

ABSTRACT

This paper presents a new object-based segmentation technique which exploits a large temporal context in order to get coherent and robust segmentation results. The segmentation process is seen as a problem of minimization of an energy function. This energy function takes into account a data attach term and spatial and temporal regularization terms. The proposed technique used to minimize this energy function is decomposed into three main steps: 1) definition of a technique for retrieving potential objects (referenced as seed extraction), 2) motion estimation for each seed, and 3) final classification performed by minimizing the energy function using a clustering-like technique. The proposed segmentation technique has been validated on real video sequences.

1. INTRODUCTION

Region-based segmentation is a very important problem for many image processing applications, such as, for example, image sequence analysis or compression. Classically, spatial or spatio-temporal criteria are used to define spatial or spatio-temporal segmentations, respectively. Many approaches have already been proposed to realize segmentation, such as motion detection with regularization constraints (e.g. Markov Random Fields) [1], region growing [2], or active contours [3, 4]. In [5], it has been shown that most of those approaches can be unified as an energetic modeling problem. Segmentation models used in the context of the famous Level Set approach are especially using this kind of model [3, 4].

In a general way, the definition of an efficient segmentation algorithm requires to clearly specify the characteristics or constraints which have to be fulfilled by the final segmentation (e.g. homogeneous texture or motion on each region, temporal consistency of the texture according to a motion model, ...). The efficiency of a segmentation algorithm will therefore depends on two main aspects: 1) the quality of its model, often defined by an energetic function and 2) the efficiency of the method used to minimize this function. In order to significantly improve the quality level reached by the state of the art segmentation methods, it seems to be necessary to introduce more complex functions than the existing ones. For that purpose, two main aspects may be considered. First, it appears that classical region-based motion representations using affine motion model usually fails to correctly represent articulated or non-rigid motions. A promising alternative consists in the introduction of a mesh representation which allows a more flexible modeling of the temporal evolutions in the images [6, 7]. A second aspect is related to the limitation generated by the classical use of only two successive images to evaluate the temporal homogeneity of the regions. The use of a longer temporal context, thanks to mesh tracking, can potentially reduce the sensitivity of the segmentation algorithm to problems such as occlusions or close motion between objects (e.g. object with low motion or not moving on a period). As a consequence, such an approach may potentially improve the stability, the robustness and the coherence of the results. Recent works have been proposed to jointly segment several images. For example, in [8] a region merging technique which takes into account all pictures of a given temporal segment is used. The merging process takes place via a spatio-temporal region adjacency graph where the vertices are merged according to the consistency criterion minimization. In this context, we propose in this paper a segmentation method based on a long term motion-based segmentation approach combined with a mesh-based tracking of the objects.

2. ENERGETIC MODEL FOR SEGMENTATION

The model used to perform the segmentation process is usually based on an energy function which should be minimized. This energy function typically contains a term which measures the adequacy of the current labeling with the observations (E^d: data attach term), and a term which takes into account the spatial or temporal context of the considered pixel (E^{rs}: regularization term). The next paragraphs describe successively the general principles of the classical approach where only two images are taken into account, and the proposed mesh-based long-term temporal approach.

2.1. Short term energetic model

When only two images are used to segment an image I_t at time t, the labeling of a pixel i to a class (or region defined for example by a motion similarity) k among K ones is obtained by minimizing a functional energy according to $P_{i,k,t}$

parameters. $P_{i,k,t}$ may be considered as the probability of pixel i at time t to belong to class k. Fuzzy techniques will consider any positive values for $P_{i,k,t}$ with the constraints that $\forall(i,t), \sum_{k=1}^{K} P_{i,k,t} = 1$ while relaxation techniques will consider that only one $P_{i,k,t}$ is non zero (i.e. 1) for (i,t). Generally the considered functional energy is:

$$E = \sum_{k=1}^{K} \sum_{i=1}^{N} \left\{ E^d_{i,k,t} + E^{rs}_{i,k,t} \right\} \qquad (1)$$

with $\left\{ \begin{array}{l} E^d_{i,k,t} = P^2_{i,k,t} \times dist(I_2(i), I_1(\Theta_k^{t_2 \to t_1}(i)))^2 \\ E^{rs}_{i,k,t} = \alpha \sum_{j \in \mathcal{V}(i)} dist(P_{i,k,t}, P_{j,k,t})^2 \end{array} \right.$,

where $dist(I_2(i), I_1(\Theta_k^{t_2 \to t_1}(i)))$ is the distance between the current image I_2 and the displaced reference one I_1 (typically quadratic error distance is used). $\Theta_k^{t_2 \to t_1}(i)$ represents the position of the i^{th} pixel in the reference frame I_1. $dist(P_{i,k,t}, P_{j,k,t})$ is the distance between the label of i compared to the class of its neighborhood $\mathcal{V}(i)$. α is a weighted coefficient which controls regularity.

The data attach term E^d stands for assigning each pixel to the best motion class by trying to minimize the quadratic error between the reference and the current image. The regularization term E^{rs} is introduced to spatially smooth the result of the classification process by penalizing label probability difference on neighbor pixels. In the case of non-fuzzy technique, this term does correspond to the Gibbs regularization term used in Markov Fields.

One limit of such technique is that a coherent labeling throughout time can not be guaranteed. Furthermore energetic model does not well represent what occurs in occlusion areas. Labeling in such areas may then be quite random such as observed in [3].

2.2. Mesh-based representation of objects

Energetic formulation of the segmentation problem as defined by Equation 1 can be linked to a model of evolution of the objects, that is $I_2(i) = I_1(\Theta_k^{t_2 \to t_1}(i)) + n_2(i)$ where $n_2(i)$ is a white Gaussian noise representing texture evolution or noise acquisition.

This model can be easily generalized to several images: $I_t(i) = M_k(\Theta_k^{t \to t_{ref}}(i)) + n_t(i)$. However in this case, motion representation has to be more complex than simple global motion. We then propose to consider active meshes to represent time evolution of an object. This model is similar to the one proposed in [7] for generalizing mosaicing techniques to non rigid objects. $M_k(i)$ then represents the value of the mosaic at any time for the k^{th} object. In this study we will consider that no texture variations occur that is $M_k(i)$ is constant along time.

2.3. Long term energetic model

In order to have a temporal stability, the previous equation is modified by considering several images and temporal consistency of the labels. To this extent, a regularization energy term E^{rt} is introduced for the temporal consistency. The energetic function becomes:

$$E = \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{i=1}^{N} \left\{ E^d_{i,k,t} + E^{rs}_{i,k,t} + E^{rt}_{i,k,t} \right\} \qquad (2)$$

with

$$\left\{ \begin{array}{lcl} E^d_{i,k,t} & = & P^2_{i,k,t} \times dist(I_t(i), M_k(\Theta_k^{t \to t_{ref}}(i)))^2 \\ E^{rs}_{i,k,t} & = & \alpha \sum_{j \in \mathcal{V}(i)} dist(P_{i,k,t}, P_{j,k,t})^2 \\ E^{rt}_{i,k,t} & = & \beta P^2_{i,k,t} \sum_{l=1}^{K} dist(P_{i,l,t}, P_{\Theta_k^{t \to t-1}(i),l,t-1})^2 \\ & & \qquad + dist(P_{i,l,t}, P_{\Theta_k^{t \to t+1}(i),l,t+1})^2 \end{array} \right.$$

Temporal probability distances used in the term $E^{rt}_{i,k,t}$ are weighted according to the probability $P_{i,k,t}$ of belonging to class k since temporal coherency depends on the class a pixel belongs to.

3. MINIMIZATION OF THE ENERGY FUNCTION

The proposed minimization technique is decomposed in three main steps: 1) definition of a technique for retrieving potential objects (seed extraction), 2) motion estimation, and 3) final classification performed by minimizing Equation 2.

3.1. Seed extraction

In order to estimate objects that are present (number and position), we first perform a motion based segmentation. Usual techniques generally search for regions that follow specific motion model (translation or affine motion,...). Although these techniques provide good results, they suffer from short time consideration. We then propose to use long term information in order to improve this segmentation step.

Given a set of T frames, motion between frames is estimated using a global mesh that is tracked along time (see Figure 1). Mesh tracking is performed with the algorithm defined in [6].

From this tracking, we can then define pixel trajectories $\{Pos(i,t)\}$ along time. We thus look for regions having motion field coherent according to affine motion model. Since affine model is a limited model, motion between not too far away frames will be considered (i.e. between t_j and $t_j + \Delta t$ with typically $\Delta t = 2$). Furthermore since result may not be perfect (occlusions are not handled with the mesh, and motion model is coarse), we will consider Fuzzy Clustering technique. We then have:

$$\min_{P_{i,k}, A_{k,t}, T_{k,t}} \left(\sum_{t=1}^{T-\Delta t} \sum_{k=1}^{K} \sum_{i=1}^{N} P^m_{i,k} \times d^2_{i,k,t} \right) \qquad (3)$$

with

$$d_{i,k,t} = \| Pos(i, t+\Delta t) - (A_{k,t}.Pos(i,t) + T_{k,t}) \|$$

where $A_{k,t}, T_{k,t}$ represent the affine motion parameters for object k between frames t and $t + \Delta t$ and m represents the fuzzy coefficient which is set to 1.6. Minimization of Equation 3 is performed iteratively in a two steps loop as in conventional fuzzy c-mean algorithms. In the first step, centroids $A_{k,t}, T_{k,t}$ are updated given $P_{i,k}$ (this is a linear regression problem wheighted by probability). In the second one, $P_{i,k}$ are updated given centroids values as follows:

$$P_{i,k} = \frac{1}{\sum_{l=1}^{K} \left(\frac{\sum_{t=1}^{T-\Delta t} d_{i,k,t}^2}{\sum_{t=1}^{T-\Delta t} d_{i,l,t}^2} \right)^{\frac{1}{m-1}}} \quad (4)$$

Pixels having high probability of affectation to a class are selected in order to define the seed of the various objects (see Figure 2). In order to define motion of these objects, we then put a mesh on each object and track their seed along time using hierarchical object mesh tracking technique defined in [6]. This hierarchical technique especially allows for spreading the motion all over the image in a consistent manner.

3.2. Resolution method

In order to estimate the segmentation, a clustering-like technique is used. For this purpose, the energy function defined in Equation 2 has to be modified as follows:

$$E = \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{i=1}^{N} \underbrace{\left\{ E_{i,k,t}^{d} + E_{i,k,t}^{rs} + {E_{i,k,t}^{rt}}' \right\}}_{E_{i,k,t}} \quad (5)$$

where

$${E_{i,k,t}^{rt}}' = \beta Q_{i,k,t}^2 \times dP_{i,k,t}^2 + \gamma [P_{i,k,t} - Q_{i,k,t}]^2$$

with

$$dP_{i,k,t}^2 = \sum_{l=1}^{K} \begin{bmatrix} [P_{i,l,t} - P_{\Theta_k^{t \to t-1}(i),l,t-1}]^2 \\ + \\ [P_{i,l,t} - P_{\Theta_k^{t \to t+1}(i),l,t+1}]^2 \end{bmatrix}$$

Probabilities $Q_{i,k,t}$ are introduced for keeping a second degree equation, while ensuring temporal continuity along valid object trajectories. $\gamma[P_{i,k,t} - Q_{i,k,t}]^2$ term is introduced to guarantee that valid object trajectories are selected accordingly to observed affectation probabilities $P_{i,k,t}$.

Minimization of Equation 5 is performed iteratively with a three steps loop in which M_k, $P_{i,k,t}$ and $Q_{i,k,t}$ are successively updated knowing the two other ones.

Update of the Mosaic images. The minimization of E according to the mosaic image M_k leads to:

$$M_k(i) = \frac{\sum_{t=1}^{T} P^2_{\Theta_k^{t_{ref} \to t}(i),k,t} I_t(\Theta_k^{t_{ref} \to t}(i))}{\sum_{t=1}^{T} P^2_{\Theta_k^{t_{ref} \to t}(i),k,t}}$$

Update of the probabilistic terms. $P_{i,k,t}$ updating is performed on sets of non connected pixels (such is the case for classical Besag's Sets [9]) and by minimizing E. Since the probabilities $P_{i,k,t}$ are constrained (i.e. $\sum_{k=1}^{K} P_{i,k,t} = 1$), we rather consider the Lagrangian functional:

$$E_\lambda = \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \sum_{k=1}^{K} E_{i,k,t} + \lambda_{i,t}(1 - \sum_{k=1}^{K} P_{i,k,t}) \right\}$$

Leading to zero the derivatives of E_λ relatively to $P_{i,k,t}$ and setting $\lambda_{i,t}$ so that $\forall i, t, \sum_{k=1}^{K} P_{i,k,t} = 1$, we obtain the following updating formulation for $P_{i,k,t}$:

$$P_{i,k,t} = \frac{\sum_{l=1}^{K} \frac{\alpha' \widehat{P}_{i,k,t} + dI_{i,l,t}^2 \widehat{P}_{i,l,t}}{\alpha' + dI_{i,l,t}^2}}{\sum_{l=1}^{K} \frac{\alpha' + dI_{i,k,t}^2}{\alpha' + dI_{i,l,t}^2}} \quad (6)$$

with

$$\begin{cases} dI_{i,k,t}^2 & = & [I_t(i) - M_k(\Theta_k^{t \to t_{ref}}(i))]^2 \\ \alpha' & = & \alpha \sum_{j \in \mathcal{V}(i)} 1 + \gamma + 2\beta \sum_{l=1}^{K} Q_{i,l,t}^2 \\ \alpha' \widehat{P}_{i,k,t} & = & \alpha \sum_{j \in \mathcal{V}(i)} P_{j,k,t} + \gamma Q_{i,k,t} \\ & & + \beta \sum_{l=1}^{K} Q_{i,l,t}^2 \times \begin{bmatrix} P_{\Theta_l^{t \to t-1}(i),k,t-1} \\ + \\ P_{\Theta_l^{t \to t+1}(i),k,t+1} \end{bmatrix} \end{cases}$$

Similarly, in the last step $Q_{i,k,t}$ updating formulation is:

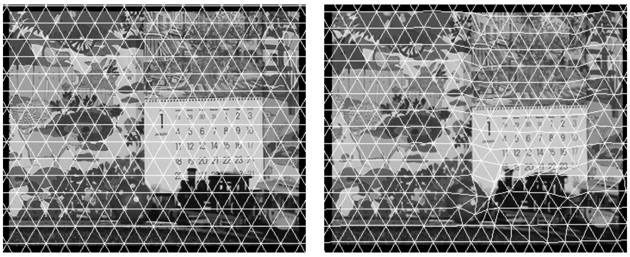
$$Q_{i,k,t} = \frac{\sum_{l=1}^{K} \frac{\gamma P_{i,k,t} + \beta dP_{i,l,t}^2 P_{i,l,t}}{\gamma + \beta dP_{i,l,t}^2}}{\sum_{l=1}^{K} \frac{\gamma + \beta dP_{i,k,t}^2}{\gamma + \beta dP_{i,l,t}^2}} \quad (7)$$

Initialization is made considering a higher probability for the terms P and Q where each seed are defined (i.e. $\forall i \in \{$seed k at time $t\}$, $P_{i,k,t} = Q_{i,k,t} = 0.6$). Moreover, another cluster is added which is the "reject cluster": $\bar{k}$. Its aim is to reject pixels which are not coherent with proposed models. The probability is computed with Equation 6, forgetting the temporal constraints. The distance $dI_{i,\bar{k},t}^2$ for this reject cluster is experimentaly set to 100.

4. RESULTS

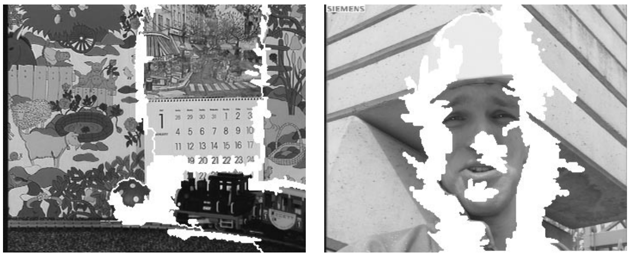
Experiments have been performed on *Mobile&Calendar* and *Foreman* sequences. Segmentation has been performed on sets of 10 frames. Figure 1 shows mesh tracking along time for *Mobile&Calendar* sequence From this tracking, as explained in section 3.1, fuzzy c-mean algorithm enables to find object seeds (see Figure 2). Motion is then estimated for the dominant seeds. In order to illustrate the segmentation technique, first computed mosaics are presented on Figure 3. It can be observed that for pixels belonging to the correct object, texture is preserved while for the others, texture gets blurred.

Finally 3D segmentation is performed with the clustering technique proposed in this paper (α, β and γ being set to 1000). Figure 4 shows results obtained after 40 iterations of the clustering technique. Objects are globally well defined with a good spatio-temporal consistency. However areas with uniform texture such as the bottom of the calendar may not be well segmented since motion does not permit to have a good discrimination between the proposed motion models.



(a) frame 0 (b)frame 10

Fig. 1. Long term mesh tracking on *Mobile&Calendar* sequence.



(a) 4 seeds (b) 2 seeds

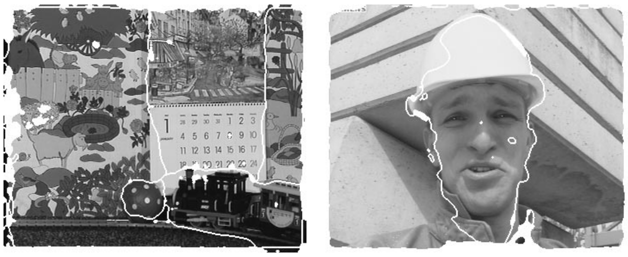
Fig. 2. Seeds extraction results based on motion clustering (white areas correspond to non-classified pixels)

(a) train's initial mosaic (b) background's initial mosaic

Fig. 3. Initial mosaics obtained with objects motion models.

5. CONCLUSION

We have presented in this paper a new segmentation technique for non rigid objects based on long term temporal consistency. This technique is based on a mesh-based representation of the objects and on the modeling of the segmentation problem as an energetic function minimization. First results show a good quality segmentation with a good spatio-temporal consistency. Future works will focus on the minimizing technique; motion refinement according to obtained segmentations, introduction of spatial constraints (adequation of the segmentation with a spatial segmentation), introduction of a multi-resolution scheme.



(a) *Mobile&Calendar* (b) *Foreman*

Fig. 4. Final segmentation

6. REFERENCES

[1] A. Caplier, L. Bonnaud, and J.M. Chassery, "Robust fast extraction of video objects combining frame differences and adaptice reference image," *ICIP International Conference on Image Processing*, 2001.

[2] R. Castagno, T. Ebrahimi, and M. Kunt, "Video segmentation based on multiple features for interactive multimedia applications," *TCSVT Transactions on Circuits and Systems for Video Technology*, 1998.

[3] A. R. Mansouri and J. Konrad, "Multiple motion segmentation with level sets," *TCSVT Transactions on Circuits and Systems for Video Technology*, 2000.

[4] S. Jehan-Besson M. Barlaud G. Aubert, "Video object segmentation using eulerian region-based active contours," *ICCV International Conference on Computer Vision*, 2001.

[5] S. C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing, and bayes/mdl for multi-band image segmentation," *TPAMI Transactions on Pattern Analysis and Machine Intelligence*, 1996.

[6] G. Marquant, S. Pateux, and C. Labit, "Mesh and "crack lines": Application to object-based motion estimation and higher scalability," *ICIP International Conference on Image Processing*, 2000.

[7] S. Pateux and G. Marquant, "Object mosaicking via meshes and crack-lines technique. application to low bit-rate video coding," *PCS Picture Coding Symposium*, 2001.

[8] B. Parker and J. Magarey, "Three-dimensional video segmentation using variational method," *ICIP International Conference on Image Processing*, 2001.

[9] J. Besag, "On the statistical analysis of dirty pictures (with discussion)," *Journal of the Royal Statistical Society Series B*, vol. 48, pp. 259–302, 1986.