

ACTIVE MESH FOR VIDEO SEGMENTATION AND OBJECTS TRACKING

Sébastien Valette, Isabelle Magnin and Rémy Prost, member, IEEE

CREATIS, CNRS Research Unit (UMR 5515) and affiliated to INSERM, INSA, Villeurbanne, France
E-mail: {sebastien.valette, isabelle.magnin, remy.prost}@creatis.insa-lyon.fr

ABSTRACT

The new MPEG-4 standard will provide superior services for the user, as it introduces the video objects concept. However, there is no generic segmentation technique able to provide the segmentation maps for such a coding system. In this paper, we propose a novel mesh-based video segmentation and objects tracking algorithm with both robust motion estimation and modeling of motion discontinuities. In addition, occlusions and uncovered regions are well managed. This allows the mesh deformation without the need to process remeshing in motion occlusion regions. The spatial properties of each frame are considered to make the mesh edges fit the image contents and a temporal smoothness constraint is also implemented. The proposed algorithm is able to segment multiple objects at the same time. Some experimental results are shown.

1. INTRODUCTION

New audio-visual services will be based on MPEG-4 and MPEG-7 standards which consists in video objects (VO) handling. This requires a video object plane (VOP) for the description of each object. Active meshes historically designed for low bitrate video compression are good candidates for this task [1-6]. Some previous work have combined spatial segmentation and motion estimation in order to track the VOs along the sequence: the frame are spatially segmented in small patches. Afterwards the motion is estimated for each patch, in order to project the first segmentation map throughout the sequence [7] [8]. In most cases, motion estimation is computed using differential techniques, based on the minimization of the motion compensation error between two consecutive frames for each patch. Then we can point out an issue : such kind of motion estimation gives good results when applied on highly textured regions, but becomes less accurate in spatially homogenous ones. In sharp contrast with these techniques, the proposed algorithm computes robust motion estimation in spatially heterogeneous cells.

In this paper, we assume that a segmentation map for the first frame of the sequence is manually or semi-automatically build.

In the second section of this paper, we introduce the proposed mesh design algorithm. In section 3, a new energy functional is described, which allows meshes to fit the image contents. In section 4, we explain the proposed splitting criterion, needed for the mesh refinement. Section 5 depicts some experimental results, and the conclusion follows.

2. MESH DESIGNING

2.1. Related works

Two main approaches come to generate a content-based mesh structure for a given image of a sequence. In [5], the nodes are first extracted from the image characteristics, such as the Displaced Frame Difference (DFD) and the spatial gradient. Afterwards, a mesh is built with these nodes, using the constrained Delaunay triangulation. In [1] and [2], the mesh is first designed as a regular mesh and the nodes are displaced inside the frame, by minimizing an objective function, so as it fits the image contents. The mesh can be refined in a quadtree selective splitting step, when more details are necessary.

As previous work aimed at providing good motion compensation models for low bitrate video compression, the motion is generally modeled by the displacement of the mesh nodes, so that the amount of motion information to be transmitted stays small. Unfortunately, this approach enforces the estimated motion field to be continuous, and performs poorly in occlusion regions, where motion discontinuities occur, such as 'Background To be Covered' (BTBC) and 'Uncovered Background' (UB) regions. Note that an occlusion adaptive mesh designing scheme is proposed in [5], where remeshing is applied where occlusions appear, allowing the mesh to track objects along the video sequence.

2.2. Our proposal

In sharp contrast with [5], we propose a novel mesh designing scheme, where the error introduced by motion discontinuities for motion estimation can be greatly reduced as in [9], and where no remeshing is needed for BTBC and UB.

We chose a triangular hierarchical mesh designing method, where the mesh is alternatively deformed and refined, in a quadtree splitting scheme. The deformation makes the mesh fit the image contents, by minimizing an objective function described in section 3. Figure 1 depicts some steps of mesh designing for the first frame, starting with a uniform and coarse mesh.

As the mesh is deformed along the video sequence, remeshing is sometimes necessary in order to keep some cells from becoming too small or degenerated.

After the mesh has been built for the first frame, it is considered as the initial mesh for the following frame, and the deformation, splitting and remeshing steps are performed again.

3. OBJECTIVE FUNCTION

3.1. Global energy

The mesh deformation is the most important step of the mesh generation process. We designed our algorithm considering three observations:

- Each triangular cell has to contain a region with homogenous dense motion field so that no cell contains two different VO.
- At the current frame t , the segmentation map at frame $t-1$ has already been built. Hence a backward projection of each cell from frame t to frame $t-1$ is a good indication on the quality of the current segmentation versus the previous one.
- The mesh edges placed on VOs boundaries should lay on high gradient regions.

As a consequence, in our algorithm, the mesh is deformed at each frame by minimizing the following energy E :

$$E(t) = E_m(t) + \alpha.E_b(t) + \beta.E_s(t)$$

$E_m(t)$, $E_b(t)$ and $E_s(t)$ are respectively the motion energy, the backward projection energy and the spatial boundaries energy, respectively. The constants α and β are positive balancing parameters between these three terms. Each of the three energy terms will be described in detail in the next sections. The minimization is processed by iterative refinements.

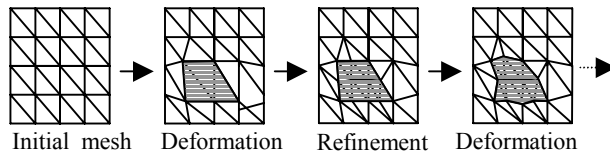


Figure 1 : mesh design scheme overview

3.2. Motion energy

The motion energy is defined as :

$$E_m(t) = \frac{\sum_i \left(\sum_{p \in c_i} [I(\tau_i(p), t + d_i) - I(p, t)]^2 \right)}{\sum_i a_i}$$

where:

- $I(p, t)$ is the intensity of the pixel p at the location (x, y) in the current frame at time t .
- a_i is the area of the cell c_i .
- τ_i can be a forward (τ_i^+) or backward (τ_i^-) transformation.
- d_i is the direction (forward or backward) of the considered cell motion.

Finally, $E_m(t)$ is the normalized quadratic motion compensation error for the frame t , with the distinctiveness that for each cell c_i , the motion is estimated with both backward and forward methods, in order to lower the error caused during motion estimation by BTBC and UB regions: a BTBC region is better suited for backwards motion estimation and a UB region will only be well estimated by forward motion estimation.

The comparison between the two resulting motion compensation errors will give the direction d_i of the considered cell motion :

- if the smallest error is obtained by backward estimation, then $d_i = -1$ and $\tau_i(p) = \tau_i^-(p)$.
- otherwise $d_i = 1$ and $\tau_i(p) = \tau_i^+(p)$.

We made our experiments with two motion models. These were chosen for their ability to handle motion discontinuities:

- A constant motion model (two parameters per cell), where discontinuities are allowed between each cell of the mesh
- A Constrained affine model (six parameters per cell), where motion discontinuities are allowed only between the different VOs.

3.3 Backward projection energy

The backward projection energy is described as follows:

$$E_b(t) = \frac{\sum_i m_i}{\sum_i a_i}$$

where m_i is the number of misclassified pixels for the backward projection of the cell c_i , computed as follows: Each cell c_i is backward projected to the segmentation map constructed in the frame $t-1$ (using the backward transformation τ_i^- constructed during backwards motion estimation) and is associated to a VO in the previous frame. Afterwards, the number of pixels which were projected outside the associated defines m_i .

As an example, the cell c_1 in figure 2 is associated to the object 1 in frame $t-1$, and m_1 is equal to the number of pixels in the grey colored region R.

The $E_b(t)$ energy adds a temporal smoothness constraint on the segmentation. It also counterbalances the motion energy in homogenous regions of the image, where the motion compensation error is inaccurate.

3.4. Spatial boundaries energy

Consider :

$$E_s(t) = \frac{\sum_i l(e_i) \cdot r(e_i)}{\sum_i \left(r(e_i) \cdot \sum_{p \in e_i} grad(p, t)^2 \right)}$$

where e_j is an edge of the mesh, and l_j its length.

$E_s(t)$ is an active contour based energy, constraining the outline of each object to be located on high gradient parts of the image. Considering an edge e_i , if it is inside an object (i.e. the two triangular cells sharing it are associated to the same object) then $r_j = 0$, otherwise $r_j = 1$.

3.5. Balancing parameters

The choice of the two balancing parameters α and β is very important, and can change depending on the sequence to be processed. For a sequence which contains a lot of motion information, the motion energy should be the most important term in the objective function.

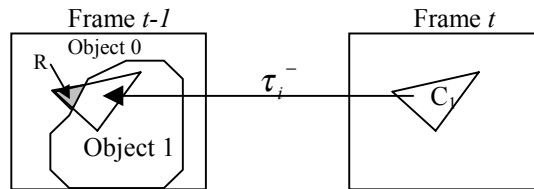


Figure 2 : backward projection energy processing.

In the other hand, a video sequence with poor motion information should favor the backwards projection energy and the spatial boundaries energy.

4. CRITERION FOR CELL SPLITTING:

After the mesh is deformed, each cell of the mesh has to pass the split test: for a given cell c_i , if $\frac{m_i}{a_i}$ is greater

than a given threshold S then the cell has to be subdivided. As an example, in figure 2, the cell c_1 has been associated to the VO #1, but a part of it lays on the VO #0 too, as shown in figure 2, $\frac{m_1}{a_1}$ should be greater

than S, resulting in the subdivision of this cell. The difficulty of this approach is the choice of the threshold. However, our experiments have shown that it can be determined by successive tries on a training sequence. In practice, the method is quite robust to this choice.

5. EXPERIMENTAL RESULTS

Figure 3 depicts the results obtained by applying our algorithm on the "van" sequence. Although our method needs an initial segmentation, it can start creating a segmentation automatically, by detecting the moving regions of the first processed frame. Mathematically this results, for the first frame, in using only the motion energy for the mesh deformation, and using a splitting criterion based on the motion compensation error. The automatically detected moving regions and the associated mesh for the frame #100 are shown in figure 3-a). Unfortunately this automatic segmentation does not satisfy our goal, since the moving van and the car following it are considered as part of the same object, and the objects boundaries are not always well located.

As a consequence, some corrections have to be performed on this first segmentation : objects redefinition (to split the detected moving object into the moving van and the moving car) and boundaries adjustment (to make the mesh fit the objects boundaries. Figure 3-b) shows the modified mesh and the associated segmentation, which now satisfy our aim.

ACKNOWLEDGEMENTS :

This work was supported in part by the Ministère de l'Éducation Nationale, de la Recherche et de la Technologie, Réseau National de Recherche en Télécommunications (RNRT), OSIAM project.

This work is in the scope of the scientific topics of the PRC-GDR ISIS research group of the French National Center for Scientific Research (CNRS).

REFERENCES

[1] Y. Wang and O. Lee, Active mesh – a feature seeking and tracking image video sequence representation scheme, *IEEE Trans. Image Processing*, vol 3, pp 610-624, september 1994.

[2] Y. Wang and O. Lee, Use of two-dimensional deformable mesh structures for video coding, part II – the analysis problem and a region-based coder employing an active mesh representation, *IEEE Trans. on Circuits and Systems for Video Technol.*, vol. 6, december 1996, pp. 647-659.

[3] L. Huang and C.-Y. Hsu, “A new motion compensation method for image sequence coding using hierarchical grid interpolation”, *IEEE Trans. on Circuits and Systems for Video Technol.*, vol. 4, pp 72-85, 1994.

[4] Y. Altunbasak and A. M. Tekalp, Closed-form connectivity-preserving solutions for motion compensation using 2-D meshes, *IEEE Trans. Image Processing*, vol. 6, no 9, pp. 1255-1269, September 1997.

[5] Y. Altunbasak and A. M. Tekalp, Occlusion-adaptive, content-based mesh design and forward tracking, *IEEE Trans. Image Processing*, vol. 6, no 9, pp. 1270-1280, September 1997.

[6] M. H. Gökçetekin, M. D. Harmanci, I. Celasun and A. Murat Tekalp, Mesh based segmentation and update for object based video, *proceedings of ICIP 2000*, Vancouver, Canada, October 2000

[7] G. Gu and M.C. Lee, Semantic video object tracking using region-based classification. *IEEE ICIP'98*, pp 643-647, Chicago, USA, 1998

[8] S. Pateux, Tracking of video objects using a backward projection technique, *proceedings of VCIP'2000* (Visual Communication and Image Processing). Perth, Australia. June 2000.

[9] A. Chretien-Planat, Estimation de mouvement par maillage actif multiechelle avec prise en compte des discontinuités : Application à l'imagerie cardiaque en Résonance Magnétique, PhD Thesis, no 99ISAL0011, INSA, Lyon, France, 1999.

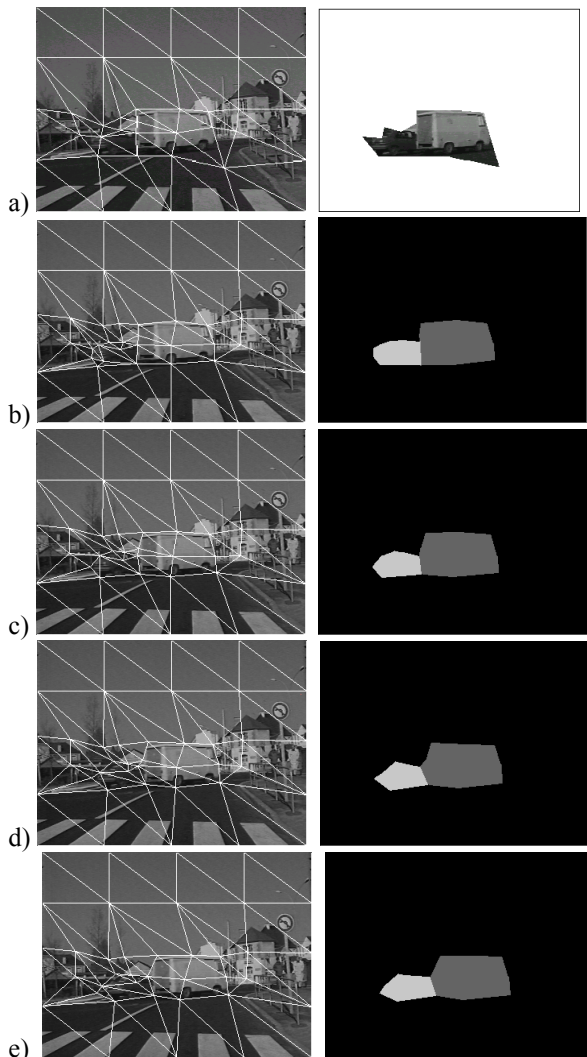


Figure 3 : results on the “van” sequence

Afterwards, the full tracking algorithm can be launched, with the initial mesh and the initial segmentation. Figures 3-c) to 3-e) show the automatically created meshes and segmentations for the following frames of the sequence.

6. CONCLUSION

In this paper we described a novel mesh designing scheme, allowing motion discontinuities, occlusions and uncovering areas, allowing the segmentation of video objects for MPEG 4 video. Experimental results prove the efficiency of the proposed algorithm.