PERGAMON

Pattern Recognition 36 (2003) 1445–1447

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and brief communication

Spatiotemporal video segmentation and motion estimation through irregular pyramids

G. Valencia*, J.A. Rodríguez, C. Urdiales, A. Bandera, F. Sandoval

Dpto. Tecnología Electrónica, ETSI Telecomunicación, Universidad de Málaga, Campus de Teatinos, 29071 Málaga, Spain

Received 16 August 2002; accepted 29 August 2002

Abstract

This paper presents a new spatiotemporal segmentation technique for video sequences. It relies on building irregular pyramids based on its homogeneity over consecutive frames. Pyramids are interlinked to keep a relationship between the regions in the frames. Its performance is good in real-world conditions because it does not depend on image constrains.

Keywords: Spatiotemporal segmentation; Hierarchical structures; Irregular pyramids; Motion estimation

1. Introduction

Segmentation consists of dividing a scene into a set of regions which are homogeneous according to some criteria. Most computer vision applications aim at obtaining a set of meaningful objects in a scene. However, objects are usually not characterized by an homogeneous intensity or color. Hence, segmentation algorithms based on these features do not produce meaningful partitions. To overcome this problem, motion information has been recently included in many segmentation techniques. In these cases, algorithms work with video sequences. These segmentation techniques can be roughly divided into two groups: joint motion estimation and segmentation, and spatiotemporal segmentation. The first group rely on estimating motion at pixel level and segmenting the scene according to the results. The second group combines spatial and temporal information so that objects moving in a coherent way can be separated. Barron et al. [1] present an excellent review of the pros and cons of these motion estimation methods. Basically, the main problem of methods working at pixel level is that they are quite sensitive to noise and illumination changes. Spatiotemporal segmentation techniques have been reported to be computationally expensive. Hence, some authors use hierarchical structures to enhance their processing times [2,3]. They estimate motion by using classic 2D spatiotemporal segmentation procedures at coarse levels. Then, they propagate their results to higher resolution levels by correcting or predicting propagation errors. To avoid errors derived from classic procedures, the authors proposed a new hierarchical segmentation method in [4]. However, the proposed algorithm had an important drawback: connectivity in segmented regions was not granted. This paper proposes a new structure based on irregular pyramids that overcomes this problem. The structure is presented in Section 2. Section 3 presents the proposed spatiotemporal segmentation algorithm. Tests, results and comparatives with classic algorithms are presented in Section 4.

2. Structure generation

We propose a hierarchical structure based on a linked pyramid. A linked pyramid is a graph $G(V,E)$ consisting of a set of vertices V linked by a set of edges E. We refer to the vertices as nodes and to the edges as links. The base of the pyramid is designated as level 0. Each node n in a pyramid is identified by (l,i,j) where l represents the level

* Corresponding author. Tel.: +34-952-137-153; fax: +34-952-131-447.

E-mail address: gaby@dte.uma.es (G. Valencia).

0031-3203/03/$30.00
PII: S0031-3203(02)00265-0

and (i, j) are the (x, y) coordinates within the level. For level 0, since each node is related to one pixel, all nodes are homogeneous. For consecutive resolution levels, homogeneity is set to 1 if the four nodes immediately underneath have similar colors. We associate four parameters for these homogeneous nodes:

- Homogeneity, $H(x, y, l)$, is set to 1 if the four nodes immediately underneath have similar color and their homogeneity values are equal to 1.
- Color, $C(x, y, l)$, is equal to the average of the four nodes immediately underneath.
- Area, $A(x, y, l)$, is equal to the sum of the areas of the four nodes immediately underneath.
- Parent link, $(X, Y)_{(x,y,l)}$: The values of parent link of the four cells immediately underneath are set to (x, y).

When the generation step has finished, only nodes presenting an homogeneity value equal to 1 are valid. Hence, since non-homogeneous nodes are not considered furthermore, the resulting structure is an irregular pyramid. Valid nodes are linked to homogeneous regions at the base. The higher the level a node is associated to, the larger is its linked region at the base level.

3. Spatiotemporal segmentation

To segment consecutive frames, two structures as the previously described one must be constructed over them. A structure built over frame $t-1$ is going to be referred to as $S(t-1)$ and each of its nodes is identified by $(i, j, l, t-1)$. Segmentation is achieved by a relinking top-down process consisting of the following steps:

(1) *Homogeneous nodes linking*: This step is divided in two stages:

- Nodes whose parent link values are null in $S(t-1)$ are linked to the parent of the best neighbor node in $S(t-1)$, and nodes whose parent link values are null in $S(t)$ to the parent of the best neighbor node in $S(t)$.
- Each node in $S(t-1)$ tries to link to the parent of the best neighbor node in $S(t)$. Hence, nodes from different structures are linked preserving homogeneous regions.

At the end of this step, each node $(x, y, l, t-1)$ can be linked to at most two parents of neighbor cells: one in $S(t-1)[(x'_p, y'_p, l+1, t-1)]$ and one in $S(t)$ $[(x''_p, y''_p, l+1, t)]$. To create a new link between node $(x, y, l, t-1)$ and a parent in $S(t-1)$, the following steps are performed:

- The color difference between neighbor nodes is calculated as $D(x_i, y_j) = |C(x, y, l, t-1) - C(x_i, y_j, l, t-1)|$, being $x-1 \leqslant x_i \leqslant x+1$ and $y-1 \leqslant y_j \leqslant y+1$.
- All nodes whose $D(x_i, y_j) < Dist\ Max$ are considered brother nodes. *Dist Max* is a threshold that fixes the maximum dispersion of the regions at the base.
- The node presenting the minimum color difference is the best brother node and node $(x, y, l, t-1)$ is linked to its parent.

To create new links between nodes in $S(t)$ and parents in $S(t)$, and nodes in $S(t-1)$ with parents in $S(t)$, similar steps are accomplished.

(2) *Homogeneous nodes fusion*: This step is divided in two stages:

- Nodes in $S(t-1)$ are fused to brother nodes in $S(t-1)$,
  and nodes in $S(t)$ are fused to brother nodes in $S(t)$.
- Nodes in $S(t-1)$ are fused to brother nodes in $S(t)$.

To fuse two nodes in $S(t-1)$, $(x_1, y_1, l, t-1)$ and $(x_2, y_2, l, t-1)$, any of the following conditions must be true:

- $(X, Y)_{(x_1,y_1,l,t-1)} = (X, Y)_{(x_2,y_2,l,t-1)}$.
- $\{(X, Y)_{(x_1,y_1,l,t-1)} = NULL\ \&\ \{|(C(x_1, y_1, l, t-1) - C(X, Y)_{(x_2,y_2,l,t-1)})| < Dist\ Max\}$.
- $\{(X, Y)_{(x_2,y_2,l,t-1)} = NULL\ \&\ \{|(C(x_2, y_2, l, t-1) - C(X, Y)_{(x_1,y_1,l,t-1)})| < Dist\ Max\}$.
- $\{(X, Y)_{(x_1,y_1,l)} = NULL\ \&\ (X, Y)_{(x_2,y_2,l)} = NULL\}\ \&$ $|C(x_1, y_1, l, t-1) - C(x_2, y_2, l, t-1)| < Dist\ Max$.

Nodes in $S(t)$ are fused with nodes in the same structure in the same way. Finally, if the node $(x, y, l, t-1)$ is linked to two parents, one in $S(t-1)$ $((x'_p, y'_p, l+1, t-1))$ and one in $S(t)$ $((x''_p, y''_p, l+1, t))$ both parents are identified as the same region.

When this algorithm is finished, any valid node at $S(t-1)$ is linked to an homogeneous pixel region at frame $t-1$, but also at an homogeneous region at frame t. Since both regions belong to the same moving object, the displacement of such a region can be calculated as the displacement of their centroids. The main advantages of the proposed algorithm are its low computational time and it depends only on threshold *Dist Max*.

4. Experiments and results

The proposed algorithm provides motion estimation and segmentation as a result of the combined stabilization technique. Fig. 1 shows the result of applying different motion estimation and segmentation techniques to a video sequence captured with a moving camera. The scene presents a set of office gadgets over an homogeneous background, and the different method applied are: the Horn and Schunk method

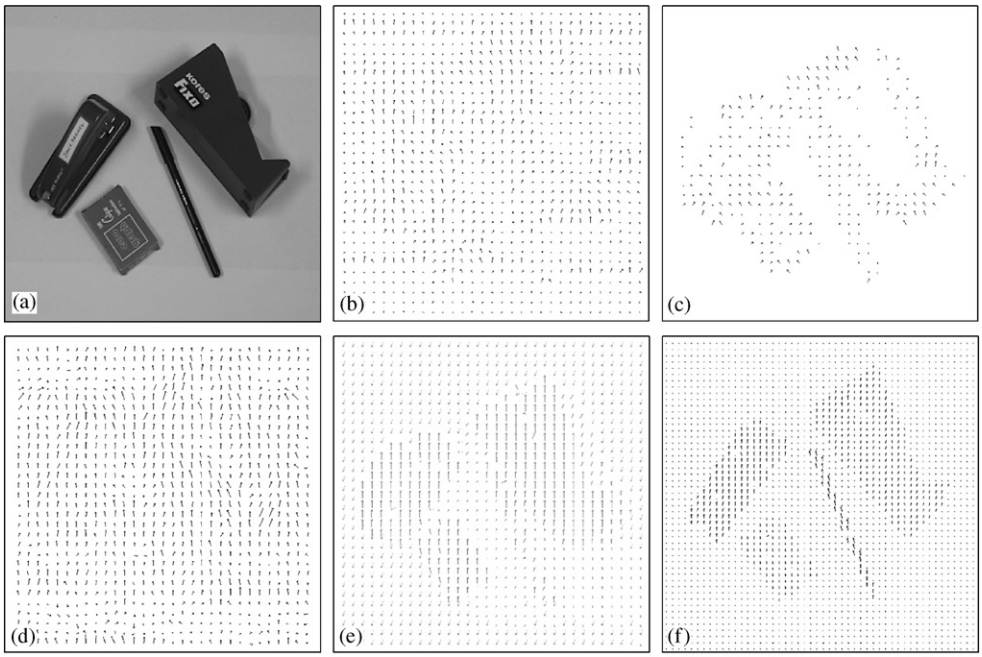

Fig. 1. Motion vector fields at two consecutive frames by using different methods: (a) original image; (b) the Horn and Schunk method; (c) the Lucas & Kanade method; (d) the Anandan method; (e) the adaptively linked pyramidal method; and (f) the irregular pyramids based method.

(Fig. 1b); the Lucas and Kanade method (Fig. 1c); the Anandan method (Fig. 1d); the adaptively linked pyramidal method (Fig. 1e); and the new irregular pyramids based method (Fig. 1f).

The expected flow field is a set of vectors for each gadget, presenting the same vertical and horizontal displacement, according to the camera movement. The differential methods (Horn and Schunk, and Lucas and Kanade) does not work properly due to the homogeneity of great parts of the scene. Anandan method presents a large amount of noise, but it can be observed that the overall estimation of the flow field tends to represent the camera movement. The first proposed method works better than the former ones, but presents some erroneous motion vectors, more visible in shadowed regions and for the background. At last, the new proposed method defines perfectly the outline of the gadgets and its motion vectors.

Also, it must be noted that while processing times for the classic algorithms of Anandan, Horn and Schunk, and Lucas and Kanade took 206, 30, and 10 s, the proposed pyramidal methods needed only 2.9 and 1.5 s, respectively, for the same video sequence and PC.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Technology and FEDER funds, Project No. TIC2001-1758.

References

[1] J. Barron, D.J. Fleet, S.S. Beauchemin, Systems and experiment performance of optical flow techniques, Int. J. Comput. Vision 12 (1) (1994) 43–77.
[2] F. Luthon, A. Caplier, M. Lievin, Spatiotemporal MRF approach to video segmentation: application to motion detection and lip segmentation, Signal Process. 76 (1999) 61–80.
[3] M. Mahzoun, J. Kim, S. Sauazaki, K.O. Tamura, A scaled multigrid optical flow algorithm based on the least RMS error between real and estimated second images, Pattern Recognition 32 (1999) 657–670.
[4] J.A. Rodríguez, C. Urdiales, A. Bandera, F. Sandoval, A multiresolution spatiotemporal motion segmentation technique for video sequences based on pyramidal structures, Pattern Recognition Lett. 23 (2002) 1761–1769.