Rapid and Brief Communication

Color-based video segmentation using interlinked irregular pyramids

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

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

Received 21 May 2003; accepted 10 June 2003

Abstract

This paper presents a new 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. Virtual nodes are considered to improve matching between low resolution levels of the pyramids. Its performance is good in real-world conditions because it does not depend on image constrains.

© 2003 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Spatiotemporal segmentation; Irregular pyramids; Motion estimation

1. Introduction

Image segmentation in complex scenarios is a hard problem because entities tend not to conform homogeneous regions with respect to simple criteria. Most times, spatial information is not enough to correctly separate it into meaningful regions. Thus, many applications rely on analyzing a video sequence to acquire additional motion information. Thus, while image segmentation tries to establish spatial relationships between adjacent pixels, video segmentation also takes advantage of spatial and temporal constrains to segment a scene into meaningful regions. Video segmentation implies: (i) estimating motion; and (ii) using motion estimation to segment a frame.

Video segmentation tends to be computationally expensive. It has been reported that hierarchical structures are useful to this respect [1]. Most hierarchical video segmentation techniques rely on estimating motion by means of differential techniques at coarse levels [2]. Then, they propagate their results to higher resolution levels by correcting or predicting propagation errors. The main problem of these methods is that differential motion estimation is sensitive to noise and illumination changes. Alternatively, qualitative methods rely on spatially segmenting consecutive frames and matching the resulting regions. The authors proposed a hierarchical qualitative video segmentation method in Valencia et al. [3]. It relied on a new 3D structure consisting in an uncomplete pyramid. However, the method was sensitive to fast motion and mild occlusions.

This paper proposes a new structure based on irregular pyramids which relies on establishing relationships between the levels of the irregular pyramids corresponding to consecutive frames in a top-down way. In this new structure, spatial segmentation is not so dependent on temporal information to avoid the problems in Valencia et al. [3]. Also, to achieve a more stable and reliable spatial segmentation, we use color instead of shades of gray. We introduce a new feature in the pyramid: virtual nodes. Virtual nodes make the levels of the pyramid more resistant against changes in the scene layout and, hence, easier to match between consecutive pyramids.

2. Structure generation

We propose a variation of the linked pyramid, which 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. In our case, we need a linked pyramid $G(V, E, t)$ for each frame at instant $t$. Each node $n$ in a pyramid is identified by $(i, j, l, t)$ where $l$ represents the level, $(i, j)$ are the $(x, y)$ coordinates within the level and $t$ is associated to the frame. In a 4-to-1 linked pyramid, each level is generated by reducing the resolution of the previous one by a factor of 4. Thus, the color of a node $(i, j, l, t)$ (parent) is calculated as the average of the four nodes immediately below at level $l - 1$ (children). We associate four parameters to each node:

- Homogeneity, $H(x, y, l, t)$, is set to 1 if the four nodes immediately underneath have color differences lower than a threshold $T$ and their homogeneity values are equal to 1. Otherwise, it is set to 0.
- Color, $C(x, y, l, t)$, is equal to the average of the colors of the four cells immediately underneath for homogeneous nodes.
- Area, $A(x, y, l, t)$, is equal to the sum of the areas of the four nodes immediately underneath.
- Parent link, $(X, Y)(i, j, l, t)$. If $H(x, y, l, t)$ is equal to 1, the values of the parent link of the four cells immediately underneath are set to $(x, y)$. Otherwise, these four parent links are set to a null value.

When the generation step has finished, non-homogeneous nodes are removed from the structure and their children become orphan nodes, which are linked to homogeneous regions at the base. The higher the level an orphan node is associated to, the larger its linked region at base level is.

### 3. Hierarchical segmentation

Our proposed segmentation algorithm consists of two stages. First, a pyramid constructed over frame $t$ is spatially segmented by rearranging it level by level in a top-down way. Then, it is temporarily linked to pyramid $t - 1$ using the motion field estimated at frame $t - 1$. The resulting classes are homogeneous in time and space and motion can be extracted from the link structure.

#### 3.1. Spatial link rearrangement

This process is applied to all orphan nodes and, for any given node $(x, y, l, t)$, it consists of the following stages:

**Parent search:** Each orphan node $(x, y, l, t)$ searches for a non-orphan neighbor in a $3 \times 3$ vicinity presenting a similar color. If such a neighbor exists, the orphan node is linked to its parent. If there are several candidate parents, the node is linked to the spatially closest one. A node $(x, y, l, t)$ is linked to the parent $(x_p, y_p, l + 1, t)$ of a neighbor node $(x, y, l, t)$ if the following conditions are true:

- $H(x, y, l, t) = 1 \& H(x_p, y_p, l + 1, t) = 1$,
- $|C(x, y, l, t) - C(x_p, y_p, l + 1, t)| < T$,
- $|C(x, y, l, t) - C(x_r, y_r, l, t)| < T$.

$T$ being the same homogeneity threshold used in the structure generation step.

**Intralevel twining:** Two neighbor nodes at the same level, $(x_1, y_1, l, t)$ and $(x_2, y_2, l, t)$, are fused if the following conditions are true:

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

If two or more orphan nodes are twined, a virtual parent is generated. This parent does not belong to the original pyramid so that class connectivity can be preserved. Nevertheless, it is used to store all information in a hierarchical way. All orphan twined nodes are then linked to the same virtual parent. The virtual node yields a color value equal to the average of the color values of its sons.

#### 3.2. Temporal link rearrangement

In order to match the regions defined in consecutive pyramids, the following algorithm is performed level by level in a top-down way:

- For each orphan node $(l, i, j, t)$ not linked yet to pyramid $t - 1$, check if there is a node in the 8-neighborhood of node $(l, i + M_x, j + M_y, l - 1)$ whose color difference with $(l, i, j, t)$ is lower than the previously defined threshold $T$, where $M_x$ and $M_y$ are the estimated motion vectors for node $(l, i, j, l - 1)$ projected in the $x$ and $y$ axes, respectively. If such a node exists, establish a link between them. This step links regions defined at the same hierarchical levels. It can be noted that this step basically checks if a node defined at pyramid $t$ is located at the position it would occupy at pyramid $t - 1$ if our motion estimation is approximately correct.
- For each orphan node $(l, i, j, t)$ not linked yet to pyramid $l - 1$, check if there is a node in the 8-neighborhood of the four nodes immediately below $(l, i + M_x, j + M_y, l - 1)$ at level $l - 1$ whose color difference with $(l, i, j, t)$ is lower than the previously defined threshold $T$. If such a node exists, establish a link between them. This step links regions defined at different hierarchy levels because of motion, occlusions, deformations and, in general, any factor changing the scene layout in consecutive frames. It can be noted that this step basically searches for the potential children of a node defined at pyramid $t$ in the position it would occupy at pyramid $t - 1$ if our motion estimation is approximately correct.

Before two pyramids are linked, their nodes are spatially arranged into classes. Hence, when two nodes are
temporarily linked, their corresponding classes are matched. The rest of the nodes in matched classes are not analyzed and hence, the computational load of the process is reduced. Most regions are defined at the higher levels of the structure. This is interesting because working at lower resolution levels involves less computational load.

Fig. 1 shows an example of pyramid interlinking. When spatial linking is accomplished, nodes presenting the same color at any given level belong to the same class. It can be observed that four classes are defined in level $l$ at pyramids $t - 1$ and $t$. However only three of them appear in level $l$ at both pyramids (shirt, table and background), and can
consequently be linked. The black class defined in level $l$ at pyramid $t$ finds a match in level $l-1$ at pyramid $t-1$. It is necessary to wait until the racket class appears at pyramid $t$ in level $l-1$ to find a valid match at pyramid $t-1$.

Fig. 2 presents frame $t$ (Fig. 2a), the final segmentation results (Fig. 2b and c) and the estimated motion of the resulting classes (Fig. 2d). It can be noticed that: (i) the tennis table presents motion zero; (ii) the background presents near zero motion, but has a slight variation because of the moving objects occlusions and deformations; (iii) even though both the hand and the ball are split into two classes, their motion is similar as expected; (iv) areas smaller than a few pixels, which are marked in black in Fig. 2b, are not matched because it would be difficult to determine a correct match; those areas typically belong to boundaries; and (v) the spatio-temporal nature of the linking process makes segmentation resistant to many lighting effects because classes are forced to be consistent both in time and space.

4. Conclusions

This paper has presented a new segmentation technique based on spatial and temporal information. The proposed technique relies on using an uncomplete pyramid where nodes are related to groups of pixel presenting similar colors to achieve homogeneous color regions. These nodes are matched to nodes defined in previous pyramids by means of temporal criteria. Then, the final number of different regions is obtained by combining all available spatial and temporal information. The proposed algorithm presents the following advantages: (i) it works in a very fast way because most processing is performed in a top-down way; (ii) virtual nodes simplify the matching process because all information tends to be arranged in the same hierarchy levels for consecutive pyramids; and (iii) class connectivity is preserved and, hence, motion estimation errors due to non connected classes are eliminated.

Acknowledgements

This work has been partially supported by the Ministerio de Ciencia y Tecnologia (MCYT) and FEDER funds, project No. TIC2001-1758.

References