A FRAMEWORK FOR INTERACTIVE VIDEO SEQUENCE SEGMENTATION BASED ON MULTIPLE FEATURES

R. Castagno and T. Ebrahimi
Signal Processing Laboratory,
Swiss Federal Institute of Technology
CH–1015 Lausanne, Switzerland
e–mail: Roberto.Castagno@epfl.ch

ABSTRACT

In this paper, a scheme for interactive video segmentation is presented. A key feature of the system is the distinction between two levels of segmentation, namely Region and Object Segmentation. *Regions* are homogeneous areas of the images, which are extracted automatically by the computer. Semantically meaningful *objects* are obtained through user interaction by grouping of regions, according to application specifications. This splitting relieves the computer of a fully semantic understanding of a scene, and allows a higher level of flexibility. The extraction of regions is based on the multidimensional analysis of several image features by a spatially constrained Fuzzy C-Means algorithm. The local level of reliability of the different features is taken into account in order to adaptively weight the contribution of each feature to the segmentation process. Results about the extraction of regions as well as about the tracking of spatio-temporal objects are presented.

1 INTRODUCTION

Work presented in the literature in recent years shows a growing interest in content-based manipulation of video information, as opposed to the block or pixel-based approaches adopted in currently popular coding standards, such as MPEG-2 and H.263.

On one hand, the problem of the extraction of semantically meaningful objects in video sequences has been faced by means of completely automatic algorithms, which tend to be quite complex, require delicate fine tuning and parameter setting, and often constitute *ad-hoc* approaches to specific problems. These difficulties mainly result from the higher level image understanding that is required in order to grasp the semantic content of scenes. On the other hand, the adoption of some form of human interaction in the process is now becoming commonly accepted and can represent an added value of the method, rather than a limitation. In his book "Being Digital" [1], Nicholas Negroponte states that *"interaction is implicit in all multimedia. If the intended experiences were passive, then closed-captioned television and subtitled movies would fit the definition of video, audio and data combined"*.

In this paper, a scheme is presented in which the intervention of the user aims at easing the segmentation process by reducing the complex tasks related to semantic issues. This approach has the advantage of a higher level of robustness, as well as an increase of the flexibility of the system in view of different possible applications. In our scheme, a lower level segmentation into homogeneous regions is performed in an automatic mode by the system based on the combination of different image features. The user can interact with the process so as to obtain a higher level segmentation, resulting in the extraction of semantically meaningful objects.

2 REGIONS AND OBJECTS

The concepts of *region* and *object* are often used as synonyms in the literature related to segmentation and computer vision in general. In most cases, the two terms are used as synonyms. However, in this work we propose a distinction between the two.

We define a *region* as an area of the scene which is homogeneous according to given quantitative criteria, such as gray level, color, texture, motion or –in the most general case– a combination of them. It is important to stress that at this stage we do not require an area to have any intrinsic semantic meaning in order to be classified as region.

Our definition of *object* is in full accordance to the concept of Video Object defined in the framework of MPEG-4 [2], as an entity in a scene that a user is allowed to access (seek, browse), and manipulate (cut and paste). Unlike the above-mentioned regions, objects are strongly characterized by their semantic content, whereas they can easily lack global coherence in color, texture, and movement.

According to these definitions, and without loosing generality in most real-world situations, an object is assumed as constituted by one or more regions. The grouping of the regions into objects is dependent on a semantic interpretation of the scene, which can in turn depend on the specific application.

3 THE MULTIPLE FEATURE APPROACH

3.1 The feature vector space

From the previous discussion, semantics emerge as the key to this distinction. In order to obtain the regions, no semantic knowledge is required from the system, but a *semantic step* is needed in order to group the regions into objects. The fact that the user deals with the semantic aspect of the segmentation reduces the ill-posedness of the segmentation problem. The computer is not required to achieve any scene understanding, but can concentrate on the automatic extraction of the regions by exploiting the coherence among different features. In this aspect, the starting point of our method is similar to the one proposed in [3]. Each frame in the sequence is analyzed by a program that extracts a vector of feature values for each pixel.

We have tested approximately 20 different features, chosen among color components (such as RGB, YUV, LSH, normalized RGB and others), displacement values (the horizontal and vertical components of an optical flow), position values (the absolute x and y coordinates), and texture information. Each one of these features can be appropriately pre–processed by means of a filter chosen from a repertoire (median, morphological, low pass, etc).

The use of features that differ in range and importance poses the two-fold problem of accounting for their different ranges of variation (scaling), as well as for the different importance that has to be attributed to different features (weighting). The following sections present the proposed solutions to this problem, with the ultimate goal of appropriately defining a distance in the features vector space.

3.2 Scaling and weighting

The features that we propose to use in our segmentation scheme belong to four groups (color, motion, position and texture), each one characterized by different ranges of possible values. In order to normalize these values, the so called Mahalanobis distance is adopted, where each feature is normalized with respect to its standard deviation over the entire image.

In addition to the scaling described above, each feature needs to be weighted also according to its level of reliability in view of the segmentation.

In particular, motion estimation is known to be quite accurate –and therefore reliable– in textured areas; in uniform areas instead, the motion field is usually quite noisy. On edges, eventually, motion information tends to be inaccurate since estimation methods typically impose smoothness constraints on edges, which results in an estimation error when objects with different motions overlap.

On the other hand, spatial information (i.e. gray level and color) is reliable in uniform areas and on edges, while it tends to yield over-segmentation in textured areas. Our goal is therefore to perform the segmentation based to the most appropriate features according to a preliminary test of reliability.

In our tests, we used a classical optical flow estimation algorithm, proposed by Lucas and Kanade in [4]. This algorithm provides also an objective measurement of the local level of reliability of the motion information [5], which has been exploited in order to analyze the image and adaptively weight the different kinds of information.

Results of this analysis are shown in Fig. 1.

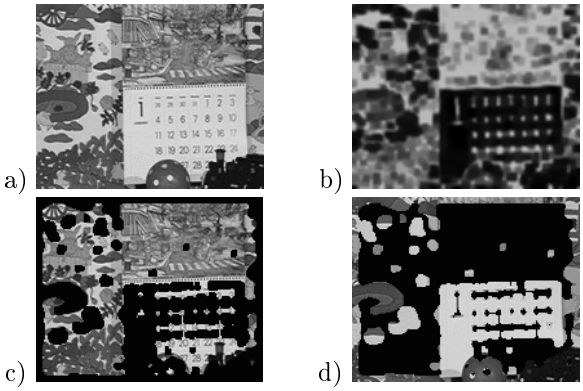

Figure 1: **a)** A frame of sequence *mobile and calendar*; **b)** The measure of reliability for motion information; **c)** Areas in the original gray level are have reliable motion estimation and will be segmented mainly according to motion information; **d)** Areas in the original gray level are have unreliable motion estimation and will be segmented mainly according to color information

Figure 1a shows an original frame of sequence *mobile and calendar*, and Fig. 1b shows the values of the measure of confidence for the motion information: a lighter gray indicates higher reliability. The two following test images (Fig. 1c and d) have been produced in order to better explain the mechanism of the estimation of reliability. In Fig. 1c , the pixels in the original gray level are those where motion information has a reliability above the average, while the remaining pixels are painted in black. Figure 1d shows the dual situation. It can be seen that textured areas are correctly identified. The algorithm will then use in each area the most appropriate features for the segmentation. For instance, the upper part of the calendar will be segmented mainly based on its motion, which is reliable and coherent, whereas a segmentation based on color would produce many small regions. Results presented in Sec. 4 demonstrate that this analysis yields efficient results when applied to the clustering process.

3.3 Constrained Fuzzy C-means

Once we have selected the features needed for the segmentation process and an appropriate distance in the vector space has been defined, a clustering is needed in order to segment the image into regions. The Fuzzy C-

Means (FCM) algorithm can be considered as a fuzzy generalization of the hard C–means algorithm [6].

Given the feature space $V = v_1, \ldots, v_N$, which represents the data set and the desired number of classes c, $2 \leq c \leq N$, the algorithm aims at finding a fuzzy partition U of the data set containing N elements:

$$U \mid u_{ik} \in [0,1] \, \forall i,k; \sum_{i=1}^{c} u_{ik} = 1 \, \forall k; \, 0 < \sum_{k=1}^{N} u_{ik} < N \, \forall i \tag{1}$$

where u_{ik} represents the degree of belongingness of feature vector $\mathbf{f}_k$ to the class i. The algorithm aims at evaluating the partition that minimizes the functional expressed by:

$$J_{FCM}(U, \mathbf{v}) = \sum_{k=1}^{N} \sum_{i=1}^{c} u_{ik}^{m} (d_{ik})^2 \tag{2}$$

where $\mathbf{v} = [\mathbf{v_1} \ldots \mathbf{v_c}]$, is the vector of the centroids corresponding to each of the classes and $m \in [1, \infty)$ is a weighting exponent that controls the amount of fuzziness. d_{ik} is the distance between the i-th centroid $\mathbf{v}_i$ and the feature vector corresponding to the k-th pixel, $\mathbf{f}_k$. The different features are weighted according to the criterion introduced in Sec. 3.1. In particular, for each pixel k we calculate a vector $\mathbf{w}_k = [w_{k1} \ldots w_{kF}]$ in which w_{kj} represents the relative weight of the j-th feature in pixel k, evaluated according to the estimate of reliability presented in Sec. 3.2. In the experiments we gave a weight of 10% to the position information (x and y coordinates), 5% to the texture information. The motion and the color information adaptively share the remaining 85% according to their reliability. In the experiments, the minimum and the maximum values are 0 and 75% for the motion information, 10% and 85% for the color features. After the introduction of the weighting factor, the distance in the feature space between the i-th centroid $\mathbf{v}_i$ and the feature vector corresponding to the k-th pixel, $\mathbf{f}_k$ is expressed as:

$$d_{ik} = \sqrt{\sum_{j=0}^{F-1} w_{kj} \frac{(f_{kj} - v_{ij})^2}{\sigma_j^2}} \tag{3}$$

where σ_j^2 is the variance of the j-th feature over the image.

The Fuzzy C-Means algorithm iterates, evaluating at each step new centroids and a new fuzzy partition, until stability is reached. For further details about the implementation of the Fuzzy C-Means algorithm, as well as for a discussion about its convergence properties, the reader is referred to [6] and [7].

It should be observed that the functional of Eq. 2 does not take into account the spatial adjacency of the data elements, which results in a lack of spatial cohesion of the resulting clusters. In order to reduce this problem, Schroeter proposed in [7] the introduction of a spatial constraint that biases the algorithm so as to encourage adjacent pixels to be assigned to the same class. The proposed modification, called Constrained Fuzzy C-Means (CFCM) aims at the minimization of the objective function:

$$J_{CFCM}(U, \mathbf{v}) = \sum_{k=1}^{N} \left(\sum_{i=1}^{c} u_{ik}^{m} (d_{ik})^2 + \sum_{h \in \eta_k} \beta \parallel \mathbf{u}_h - \mathbf{u}_k \parallel^2 \right) \tag{4}$$

where $\mathbf{u}_k = u_{1k}, \ldots, u_{ck}$, η_k is a neighborhood of pixel k and β is a parameter that regulates the strength of the spatial constraint. In our experiments, the neighborhood of pixel k consists of the four pixels in positions North, South, East and West. As it is the case for the FCM algorithm, the fuzzy partition of the data set can be obtained by iteratively updating the centroids and the degree of belongingness of each vector to the classes. In our approach, the standard FCM algorithm is used to obtain an initial segmentation, and the spatial constraint is introduced in a second round of iterations aimed at refining the result.

Once the iterations have stabilized, the algorithm is stopped and the fuzzy partition is hardened, i.e. each pixel is assigned to the class for which it shows the highest degree of belongingness. The segmentation results obtained for the current frame n are eventually used as initialization for the FCM procedure at frame $n+1$. In more detail, the matrix U for frame $n+1$ is initialized based on a motion compensated projection of the segmentation labels of frame n on frame $n+1$, and assigning initial values $u_i k$ according to :

$$u_{ik} = \begin{cases} 1 & \text{if } l_{mc}^{(n)}(k) = i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $l_{mc}^{(n)}(k)$ is the class label of pixel k in frame n after motion compensation. A set of new centroids is then evaluated based on this first estimate of the membership matrix and the process iterates until stability is reached. This mechanism of "educated initialization" allows us to obtain a good degree of stability by starting the initialization from a point which is likely to be close to a minimum, and in particular to a minimum which will show good temporal coherence with the solution obtained for the previous frames.

This tracking at the region level eventually induces a tracking at the object level: regions which are attributed to the same object by the initial user interaction are grouped together in the following frames, thus reconstructing the successive temporal instantiations of the visual objects (VOP in the MPEG–4 terminology). Tracking results are presented in Fig. 4.

4 RESULTS

The proposed method has been tested on several sequences, mainly chosen among the MPEG standard test

sequences. We propose in the following some frames taken from three typical sequences used in the tests: *table tennis*, *mobile and calendar* and *foreman*.

Figure 2a shows the results of the segmentation at the region level on a frame of sequence *table tennis*. It should be observed that —despite the fact that the hand is a highly textures area— it is grouped correctly in a single region. In fact, that area is segmented mainly based on motion information.

Figure 2b shows an output of the scheme after the replacement of the background obtained after a selection of the user. Figure 2c shows instead the results obtained by grouping all the regions characterized by motion. This result could have been obtained in theory also by a fully automatic scheme in which, after the creation of the regions, the grouping into objects is done based on a motion criterion. Figure 3 shows results of

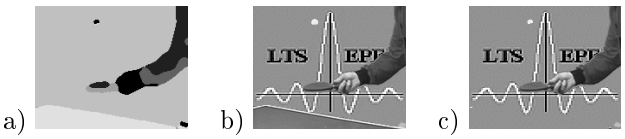

Figure 2: **a)** The regions extracted from the segmentation of sequence *table tennis*; **b)** Replacement of the background of sequence *table tennis*; **c)** Extraction of moving objects from sequence *table tennis*

object extraction from sequence *mobile and calendar*. It should be observed that —despite the high amount of texture—, the area corresponding to the calendar is labeled as a single region, because the segmentation is done based on the motion information, which is here coherent and reliable.

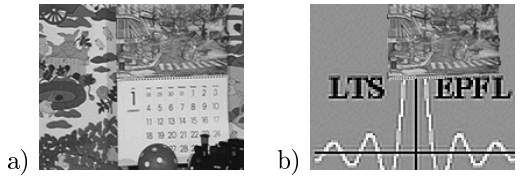

Figure 3: Segmentation results on sequence *mobile and calendar*: **a)** the original frame **b)** an extracted region

An example of the tracking results is shown in Fig. 4, for sequence *foreman*. An object has been defined by the user by grouping three regions and is tracked through time.

5 CONCLUSION

In this paper, an algorithm for the interactive segmentation of video sequences has been proposed. The scheme is based on an automatic segmentation algorithm that extracts homogeneous regions based on an adaptive combination of multiple features. The relative weights of motion and spatial information are evaluated locally based on their level of reliability. The obtained regions can

Figure 4: An object extracted from frames 2, 5, 8, 11, 14, 17 of sequence *foreman*.

be grouped into semantically meaningful objects through the interaction of the user, thus overcoming the major limitations that fully automatic methods meet when dealing with the semantic understanding of the scene. Simulation results show that this approach produces a robust segmentation, still retaining a high level of flexibility in view of a wide range of applications.
For a more complete description of the proposed algorithm, the reader is referred to [8].

References

[1] N. Negroponte. *Being Digital*. Hodder & Stoughton, London, 1995.

[2] MPEG Video and SNHC Groups. "Committee draft of MPEG–4, part 2, 14496–2". Technical Report ISO/IEC JTC/SC29/WG11/N1902, ISO/IEC, Fribourg, Switzerland, October 1997. Available on the WEB at `http://drogo.cselt.stet.it/mpeg/`.

[3] Ed. Chalom and V. M. Bove. "Segmentation of an image sequence using multi–dimensional image attributes". In *Proceedings of the International Conference on Image Processing ICIP*, Vol. 2, pp. 525–528, Lausanne, Switzerland, September 1996.

[4] B. Lucas and T. Kanade. "An iterative image registration technique with an application to stereo vision". In *Proceedings of DARPA Image Understanding Workshop*, pp. 121–130, 1981.

[5] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. "Performance of optical flow techniques". *International Journal of Computer Vision*, Vol. 12, No. 1, pp. 43–77, February 1994.

[6] J.C.Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithm*. Plenum Press, New York and London, 1981.

[7] Ph. Schroeter. *Unsupervised Two-Dimensional and Three-Dimensional Image Segmentation*. PhD thesis, Swiss Federal Institute of Technology, Lausanne, Switzerland, 1996.

[8] R. Castagno, T. Ebrahimi, and M. Kunt. "Video segmentation based on multiple features for interactive multimedia applications". *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image and Video Processing for Emerging Interactive Multimedia Services*, September 1998. To be published.