VIDEO SEGMENTATION WITH SUPERIMPOSED MOBILE MAPS OF DISTANCES

A.Viloria , J. Finat [b] , and M.Gonzalo-Tasis [b]

aviloria@lpi.tel.uva.es, [b]MoBiVA Group, Lab.2.2, Edificio I+D, Univ. of Valladolid, Spain ,jfinat@agt.uva.es, marga@infor.uva.es

Commission V, WG V/3

KEY WORDS: *Vision Sciences, Segmentation, Sequences Application, Surveying*

ABSTRACT

A real-time and reliable automatic video segmentation is one of the outstanding problems in Computer Vision with large applications to compression, transmission and motion analysis. In this paper, we show a novel approach based on the superposition of distance maps linked to centroids of mobile regions acting as attractors of homogenous regions to which different thresholds are applied. The homogeneity of each region is characterized by colour characteristics. The number of colors and the extremal values allowed for parameters corresponding to the shape of regions can be previously configured or learned through an unsupervised training. Our real-time processing does not depend on the scene complexity and it is compatible with egomotion, i.e. it is not necessary to discriminate beforehand between foreground and background. Compatibility of our segmentation algorithms with egomotion allows the design of on-line tracking and shots identification for automatic segmentation of video sequences by using a low-level topological representation, which is symbolically represented by means of a kinematic mobile graph.

1. INTRODUCTION

The increasing power of personal computers and a higher performance of algorithms allow to extend the analysis of mobile data to currently available digital libraries, including video files in different formats. Main problems relative to image processing concern to the computer implementation of segmentation and matching algorithms for mobile data. The extension of video devices requires the design of computer tools able of supporting user interaction, eventually based on a graphics interface to refine interaction. To satisfy different user's needs relative to different search procedures, it is commonly accepted that a hierarchised, hybrid and multilayered approach is required. *Hierarchies* concern to the identification of events allowing to separate units of analysis (isolated events, shots, scenes) along a video sequence. *Hybrid character* must include aspects relative to low-level image features (colour, textures, e.g.) and high-level image features (shapes, geometric primitives, e.g.). *Multilayered approach* is translated to several levels of analysis going from low-level contents retrieval in video sequences to the high level interpretation of scenes.

Following an increasing complexity order, we can consider spatial, temporal and spatio-temporal segmentation. *Spatial segmentation* is a decomposition of a view in static homogeneous regions. Homogeneity depends on the chosen threshold for meaningful characteristics (colour, texture) from the information arising from histograms.

It is difficult to give an objective evaluation of the "goodness" of a segmentation ([Pal93]). An evaluation and comparison of static segmentation to the mid of nineties can be read in [Zha97]. In this work we are more interested about some general problems concerning to mobile segmentation linked to temporal and spatio-temporal modelling. Hence, from the static viewpoint our emphasis is put on the description of individual events, i.e., (dis)apparition of larges regions besides a critical size. A recurrent problem ([Alt00]) is the automatic selection of *thresholding criteria* to identify the critical phenomena for an automatic segmentation. The introduction of metric information provides an objective criterion to select critical thresholds by adding a spatial information to the viewpoint of [Alt00].

Temporal segmentation ([Kop01]) is a mobile segmentation which is focused toward the identification of "shots". A "shot" in a video sequence means a set of image frames with similar background and continuous motion. A typical example is provided by a fixed camera in an indoor scene (underground surveillance, e.g.) or in an outdoor scene (traffic surveillance, e.g.). Hence, the analysis concerns mainly to temporal segmentation, including a kinematic information about mobile objects, eventually.

High-level dynamic segmentation concerns to the analysis of "scenes". A "scene" consists of several consecutive shots that are "semantically" correlated.

Along a "scene", the background is not necessarily the same, and includes some motion of camera, usually. Hence, the analysis concerns to *spatio-temporal segmentation*, including the estimation of kinematic characteristics of mobile objects in image and the egomotion of camera.

Segmentation techniques must be applicable to static in an accurate way and able of a reliable discrimination between ego- and external motion. We have obtained meaningful results in all of them [Vil02], with a real-time and accurate results for images of arbitrary complexity in the static case, and a fast but coarser results for mobile regions, including the on-line capability of discrimination between camera motions and external movements (human bodies in TV scenes, e.g.). In this work, we extend above results by superimposing metric information in different views which is automatically generated from the construction of distance maps linked to segmented regions. A real-time update of iconic information is possible thanks to a simple propagation model. Nevertheless, it is commonly

believed that to improve the management of complex video scenes are, it is not possible to reduce ourselves to a fully automatic segmentation tools. Hence, it is necessary the development of relations between automatic and interactive tools for video processing. Thus, the current work intends to contribute to the development of such relations by means the computer implementation of some semi-automatic processing tools.

2. REGIONS-BASED FAST SEGMENTATION AND DYNAMIC MATCHING

Fast segmentation is usually based on the extraction of regions with "similar" properties. Main issues concern to the specification of similarity notions involving to low-level patterns for image processing (histograms, colour and textures) and high-level patterns for identification and comparison of shapes. Sensitivity to brightness variations and the lack of localization (position and orientation) information are two neckbottles of a strictly colour segmentation approach. Another said, low-level patterns by themselves are difficult to manage without adding spatial information relative to their eventually mobile localization. This fact justifies our hybrid (colour-position) approach. On the other hand, the high computational complexity of kinematic models linked to simple shapes, suggests adapting some kind of symbolic representation able of supporting low- and high-level patterns.

Some advantages of symbolic representations given by adjacency graphs are: simplicity, easy updating and absorption of small changes relative to image features and shapes. Video segmentation requires to identify topological changes in a sequence of adjacency graphs. Shots are defined as discontinuities of graphs for the temporal axis, i.e., some nodes representing regions are unfolded or deleted, following birth and death usual models.

Our choice for mobile segmentation is based on an extended colour segmentation. Traditional colour segmentation identifies a typical colour for regions R_i. Furthermore, we consider the mass $\mathbf{m}_i$ and a typical shape S_i for region R_i extracted as its boundary ∂R_i. The mass $\mathbf{m}_i$ corresponds to the number of pixels contained in R_i. The boundary is the conflict locus for propagation algorithms. Typical colour arises from a homogenisation above a threshold following usual competitive propagation algorithms. We have implemented two versions of competitive propagation algorithms, which play a complementary role, which are labelled as "overflow" and expansion algorithms (see the next section for details). The extraction of contours ∂R_i is performed to an iconic level, only, i.e., without assigning any kind of mathematical primitives to each component. Anyway, we can suppose that the boundary ∂R_i is piecewise smooth. So, we have a reasonable framework for some duality questions related with the symbolic management of meaningful information.

Our symbolic approach for regions segmentation of each view is based on a graph Γ. Nodes $\mathbf{n}_i$ of the graph are supported on centroids $\mathbf{C}_i$ of regions R_i arising from a color segmentation. Two nodes $\mathbf{n}_i$ and $\mathbf{n}_j$ of the graph Γ are connected by means of an edge $\mathbf{e}_{ij}$ if and only if the regions R_i and R_j have a common component in their boundary. Our algorithm design excludes the existence of quadruple points in contours segmentation. Another said, corners can belong to two or three regions, giving us double or triple points. Each corner separates the boundary ∂R_i in two components. A T-junction generates a subdivision in the oriented component of R_i where the T-confluence is generated. Hence, the eventually increased list of corners heritates also an orientation. So, a doubly connected list (d.c.l.) is automatically generated for the management of regions, contours and corners data contained in each view, in the same way as for the linear case with a similar design of pointers. In particular, for each pair of adjacent regions R_i and R_j we count twice the common component of boundary, each one with the orientation induced by that of R_i. In the same way, each corner has an oriented weight, i.e., it appears with so many orientations as the oriented edges incident at the corner.

Centroids $\mathbf{C}_i$ of regions R_i are the sites of a Voronoi diagram, with the corresponding dual representation which supports a standard combinatorial information (Delaunay triangulation). Symbolic attributes for segmented regions R_i correspond to constant functions defined on the positively oriented region R_i (it suffices to evaluate at the centroid $\mathbf{C}_i$). Matching between different regions is easier, reliable and fast thanks to the existence of common boundaries with opposite orientations. The boundary operator assigns to each region R_i its boundary ∂R_i in a piecewise smooth way. Breaking points for smoothness correspond to oriented corners, i.e. the incidence locus of at least two different colour components. The existence of a natural orientation corresponding to all elements appearing in the d.c.l., allow to verify usual properties of boundary operators (such that $\partial^2 = 0$). Hence, we can define homology groups, which provide us information about holes, or more advanced topological properties of oriented components with homogenous properties for colour and/or textures.

If incidence conditions are preserved, nevertheless some shape changes in apparent contours, then the number of meaningful connected components is constant. Elementary topological events along a video sequence are characterized in terms of elementary transformations (grouping or splitting) of regions previously existent. For a fixed camera (with a fixed background), an elementary shot is linked to the (dis)apparition of a multibody, where a multibody is characterized as a connected tree of regions with proper motion (car, animal, human body, typically). If the camera is mobile, the discrimination between egomotion and external motion can be performed with the motion analysis of background and foreground. If the common background to several views is fixed, then there is no egomotion, and it suffices to evaluate absolute motion of mobile objects foreground. Otherwise, a finer analysis is required, and relative motion of foreground is obtained from a subtraction of the observed motion of background.

3. SPATIO-TEMPORAL PROPAGATION ALGORITHMS FOR MOBILE DATA

Each region is described as a collection of contiguous pixels with a homogenous colour. Competitive propagation algorithms provide a local homogeneity with respect to the colour. We have implemented two *spatial* propagation algorithms that are labelled as "overflow" and expansion, in correspondence with linear and rotational sweep-out techniques for each image. Competitive propagation algorithms follow simple comparison criteria for pixels linked to position and colour attributes.

In our experiments, we have avoided the use of textures due to the simplicity of processing based in colour and the presence of non-textured regions in views.

By discarding small regions below a threshold, and by using path-connected constraints regions a topological map is generated jointly with a symbolic representation given by a graph. Qualitative kinematic information is obtained from evaluating growing and decreasing phenomena of "homologue regions" along a video sequence. We develop a coarse-to-fine approach for kinematics evaluation in terms of cooperative-competitive dynamical models. Competitive models work to a microlocal (pixel) level, where small differences between parameters (specific growth rates of populations and their competition effects), are in the issue of relative advantages for survivors. Cooperative models contribute to the regions homogeneity from the local viewpoint. Along a video sequence, mobile objects are in competence for the occupancy of regions; thus, global behaviour is controlled by a competitive model with three populations which are labelled as child, parents and old. Old population concerns to the initialisation of each video sequence. The transitions between populations of regions are controlled at the intermediate parents level. Prediction concerns to the child generation depending on critical values for the allowed maximum size. The application of standard morphological operators (erosion-dilatation) and their iteration (opening-closing), simplify the identification and tracking of evolving shapes, without extracting contours.

The relative linear or angular momentum of regions gives the coarsest level for the dynamic model. The mass m_i of each homogenous region R_i is represented by the number of pixels with similar colour (modulus a threshold): a) Identify stable or inertial regions as belonging to the background (relative velocity under a threshold), b) evaluate nearness for mobile regions labelled as nodes by using adjacency graphs ,c) represent mutations (births, deaths) in terms of unfolding and collapsing nodes of the graph d) evaluate relative velocities of barycenters of regions with similar colour and localization.

Any kind of spatial propagation is based on first- or second order differences of functions evaluated at pixels. Unfortunately, first order differences are very sensitive w.r.t. illumination changes and camera motions. To obtain stable, robust and accurate results in the static case, we can use LOG operators or typical Canny's operator to avoid the dependance w.r.t. orientation and illumination. Spatio-temporal version of a laplacian is given by a Laplace-Beltrami operator. However, the high computational cost for mobile data and troubles for dynamic grouping, suggests to introduce some kind of temporal average at least for three consecutive images. So, temporal average is responsible of small delays to generate meaningful regions to be sampled, identified and tracked. Criteria for temporal average are based on mediana filters for sampled images.

Cost functions associated to regional segmentation arise from a weighted balance between a) the error tolerance at low-level and b) the maximum number of regions at high-level. Both infinitesimal and local criteria require specific thresholds that can be learned in a semiautomatic way depending on the data concentration and the critical size of regions. The most accurate results are obtained by using information arising from local histogram comparisons. Thus, the selection of meaningful thresholds can be performed from the beginning by using directly a temporal average of two local histograms: a) Threshold for error tolerance is based on the selection of local maxima in histograms corresponding to the most frequent values (medianas), and simple propagation mechanisms: If the "distance value" is below the threshold, the pixel is assigned to the current region, otherwise, a new region is created; b) Threshold for maximum number of regions can be understood as a mean average problem, which represents a variable version of k-means problem, where k represents the maximum number of regions, and each pair <position, colour> provides entries for the algorithm. This technique allows to maintain constant the costs of image processing. However, the variability of data contained in video sequences makes difficult the selection of a fixed value of k

If there is no need of a live processing, it is possible to decompose each homogenous colour region in monotone or convex parts, according to usual algorithms in Computational Geometry [Ber97]. In this case, if optimal or at least more accurate results are need, our approach is enough flexible to support additional constraints. The symbolic management of this finer decomposition is labelled as an unfolding of centroids. More accurate results linked to boundaries are obtained, but matching, indexing and contents retrieval become more cumbersome. Thus, results will not be reported here.

4. AUTOMATIC GENERATION OF DISTANCE MAPS AND OPTIMALITY

We start up with a description of an unsupervised clustering technique that is based on a competitive propagation model from centroids of homogeneous regions above a critical size. An important issue is the integration of low- and high-level image features. This integration is related with a coarse colour-shape identification (for contents retrieval) and tracking of mobile data in general spatio-temporal models. Corresponding values at coordinates are weighted according to the distribution of frequencies, to increase the relevance of colour homogeneity or shape definition.

The simplest model needs to have in account centroid position and typical colours of segmented regions, which gives us a 5-dimensional parameter space. The introduction of separating hyperplanes in this 5D space is performed in a very similar way to a generalized Voronoi diagram, but following a recursive pattern with successive subdivisions in half-spaces. So, we obtain a coarse subdivision inside the 5D parameters space that can be managed in terms of binary trees search. The spatio-temporal implementation of this algorithm produces a distance-map of contiguous regions with homogenous colour. Instead of looking at adjustable weights, we can use a feedback between colour-based grouping and shape with good properties for search/optimization processes (monotone/convex subdivisions).

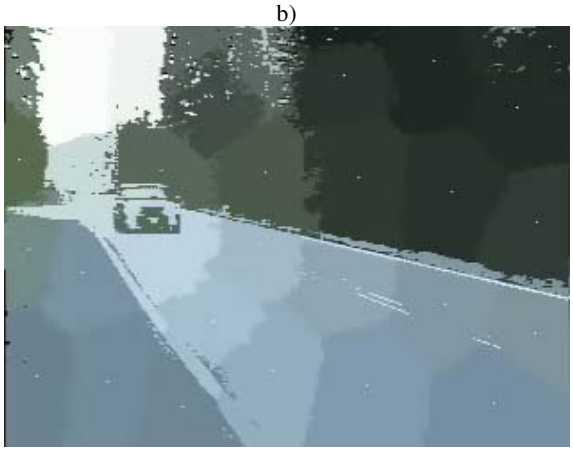
5. EXPERIMENTS

Two local factors are based in 1) the rate of the new regions created per local area unit, and 2) the size of regions. Both local factors are related between them by a hyperbolic law. So, if many regions are created in a small area, then a complex texture is found (wood or leaves in a typical outdoor scene, e.g.), and the local error threshold must be increased.

Contrarily, if we find few regions in a local area unit, the local error threshold can be lowered.

a)

b)



c)

d)

In accompanying video sequence one can see some results of our image processing method for a static camera and mobile objects. In this case, we have a fixed geometrical background with a mobile foreground. Nevertheless, the fixed character of the background, there appear some meaningful facts relative to the image segmentation which are linked to variable lightening conditions Indeed, small lightening differences are increased by concentrating available information around middle values. This sequence provides an analysis of mobile objects with easy kinematic properties relative to depth variations which can be obtained by means of an evaluation of scale parameter (by using methods appearing in [Vil02]).

Furthermore, it is possible to observe small variations in background giving us a better understanding of depth. The artificial model is linked to a discrete version of traditional human perception, where graduations are usually continuous and more slower.

Some advantages for the developed artificial system are the following ones: it is possible to give a explicit description of unfolding and regrouping processes involving to the background (including watershed effects in the ground or sky, e.g.), and, mainly, it is possible to include an artificial analogue to the role played by depth planes. In this case, depth planes are linked not to architectural elements (as it is usual in 3D Reconstruction), but to a colour segmentation with supports an optical and metric information.

The analysis for small variations of lightening and small motions in forest is similar, but including now reinforced variations of colour. These reinforced variations are supported in this case on a distance map able of reproducing small effects such as leaves frightening or more coarse variations, depending on the parameters selection. In this way, we obtain an adaptive approach to a more realistic perception, nevertheless the reduction accompanying any segmentation task.

6. CONCLUSIONS AND FUTURE WORK

In this work we have sketched some general principles for a hybrid, hierarchised and multilayered approach to video segmentation. A feedback between low- and high- level processing is revealed as a guide to combine accurate results based on local histograms with meaningful regions. Segmentation is performed in a semiautomatic way on a 5D space which is managed by using a binary tree search. Centroids of segmented regions provide sites for a generalized Voronoi diagram.

Distance maps are introduced for supporting the identification, evaluation and tracking of meaningful regions. The average between consecutive frames allows us to extend traditional accurate spatial methods to a coarse but reliable approach to the segmentation problem.

Some standing issues where more accurate results would be convenient concern to exploiting the duality sketched between the structure of boundaries and the extended Voronoi diagram, linked to centroids, the analysis of optimal clustering functions for regions matching, the comparison of coarse live results with finer results corresponding to 2^{nd} order differential spatio-temporal operators (D'Alembertian of a Gaussian) linked to mobile data. Further applications to contents retrieval and tracking in video sequences are currently in development.

7. REFERENCES

[Alt00] Y.Altunbasak: “A statistical approach to threshold selection in temporal video segmentation algorithms”, ICASS-IEEE, 2421-2424, 2000.

[Ber97] M.de Berg, M.van Kreveld, M.Overmars, O.Schwarzkopf:"Computational Geometry. Algorithms and Applications", Springer-Verlag,1997.

[Kop01] I.Koprinska and S.Carrato: “Temporal Video Segmentation: A Survey” Signal Processing Image Communication, Elsevier Science, 2001.

[Pal93 N.R.Pal and S.K.Pal: “A review on image segmentation techniques”, Pattern Recognition 26, nº 9, 1993, 1277-1294.

[Vil02] A.Viloria, J.Finat, and M Gonzalo-Tasis: “A fast self-organized iconic segmentation and grouping based in color”, V Commission, ISPRS, Corfu (Greece, Sept 2002).

[Zha97] Y.J.Zhang: “Evaluation and comparison of different segmentation algorithms”, Pattern Recognition Letters, Vol.18, 963-974, 1997.