INDEXING AND SEGMENTATION: WHAT RELATIONS ?

Henri Sanson

France Telecom R&D, 4 rue du Clos Courtel, 35512 Cesson Sévigné France,
henri.sanson@francetelecom.com

ABSTRACT

In this paper, we aim at giving a vision about the potential links between video indexing and segmentation, i.e. in what one can help the other in achieving its goals. To this end, we first review the different understandings of what is usually called indexing or segmentation. Then, from user needs and market analysis, we explain why the notion of object is inescapable in video indexing. From this, we revisit the respective roles of recognition and segmentation in providing access to semantic objects of the scene and finally deliver our intimate conviction about the winning choice at medium range.

1. INTRODUCTION

Retrieving videos of interest or relevant information therein, navigating efficiently in huge audio visual databases is an ever increasing need both for professional and private usages. As a matter of fact, more and more, people consume or use images in their every day live, for entertainment, news access, learning, communicating, promoting their products, monitoring their home, and so forth.

On the one hand, the amount of image and video data potentially accessible to the user (consumer, worker) is growing dramatically fast, making it more and more difficult to find the right content, or the information of interest conveyed in it. The issue is not only a matter of data amount, but also that of the temporal constraint inherent in the consumption of raw videos.

The right way to solve this issue is what people usually call indexing, along with the associated search procedures. By the way, what does indexing means ? In fact, there are several meanings of this term:

For librarians, this means generating high level semantic annotations that will be stored consistently with their corresponding content. Most of the time, such annotation will be manually input. This is a human-centric approach and most of the works originate from the description of textual resources, making intensive use of Natural Language Processing and Knowledge Representation, two important parts of Artificial Intelligence. The current Semantic-Web approach is quite representative of such movement.

For most image analysis specialists, such as those of the MPEG-7 movement [1], indexing means conveying in one way or another relevant descriptors of the visual content. Image scientists work on finding "signatures" able to summarize the image or video content with a high degree of relevance. The paradigm on which this approach is based is that effective retrieval capabilities can be offered on the basis of generic, universal, low level descriptors, often referred to collectively as color, shape and texture [2] and that queries are done with examples of what is searched for. If the notion of color does not pose problem of definition, it is not the case of shape and texture which still lack a clear and shared understanding. Most of the time, only global statistical features are taken into account, such as color or contour histograms.

For image sequences, shot detection and scene shots grouping provide additional descriptors of high interest. In this case, low level image analysis leads to somewhat higher level description.

For database management specialists, indexing is still associated with descriptors, but goes further by including the notion of efficient organization and search mechanisms associated to such descriptors.

Fundamentally, indexing aims at efficiently describing visual content so that a user can retrieve specific content (or information) effectively.

On the other hand, video segmentation also covers different aspects:

Temporal segmentation, in the sense of shot boundary detection, is often the first and most basic step of any automatic video indexing process. Although fully reliable temporal segmentation still deserves some research endeavor, above all as far as sophisticated shot transition effects are concerned, one may consider it a fairly straightforward and tractable issue.

Still and spatiotemporal segmentation into homogeneous regions consist in finding the areas stratifying a homogeneity criterion. The notion of homogeneity depends on the approach retained and may

be based on luminance, color or even texture consideration, but in any case, on some "surface" or "material" property of this kind.

Finally, spatiotemporal segmentation into "semantic" or at least physical objects intends to extract objects such as human beings consider them.

Temporal segmentation is undoubtedly an unavoidable stage for video indexing, which considerably relieves the processing complexity while already offering efficient access to content through delinearisation of the audio visual document. However, this does not on its own provide information about the visual content. Therefore we will not consider this aspect any longer here.

2. USER NEEDS ANALYSIS AND IMPLICATIONS ON THE INDEXING FRAMEWORK

When exploring the true needs expressed by real users, i.e. the actors of the content value chain: content owners or managers, content service providers, Internet Access Provider, Search Engines/Portals and end users, it is very difficult to find requirements that can be satisfied otherwise than with semantic descriptions of the content. Of course, in the context of focus groups, with technology-addicted people, some needs relying more on subjective or even affective queries (search for a "merry" image) are sometime expressed, but correspond very little to current practice of Internet search engines for instance, and even less of practical usage among professional people. Thus, people are looking for information about people, objects, places or events, and we have no elements to think that it could be otherwise concerning visual content. In that respect, more-like-this or query-by-example approaches do not seem to correspond to the most widespread expectation, at least when used as is.

On the other hand, there is a clear issue of saving time and/or money when performing the content annotation tasks. This issue is becoming even more dramatic with the continuous increase of visual content to deal with.

Therefore, automatic procedures enabling acceptable semantic annotation are more than expected. The automatic approach supposes to start with low level descriptors and to get up to higher level description consisting of semantic features and even interpretation of visual scenes.

Another aspect comes in favor of such an automatic recognition approach to visual content annotation: it is now a well recognized fact that there does not exist any universal annotation of visual scenes. Any semantic annotation is necessarily an interpretation of the content, therefore context and person dependent. Depending on the elements at hand, it is consequently indispensable to be able to revisit the annotation, i.e find out what had not been described in the initial or any previous indexing process.

Thus, automatic labeling of the visual content is an important issue for video indexing. This is clearly a Visual Recognition task.

A first level of labeling can be global to the image or the video shot and attribute a scene category: indoor/outdoor, city/country, ... for which global statistical descriptor can suffice. For more detailed annotation, Object Recognition is a very important task.

The question now, which has been a dilemma for long, is to determine what object segmentation and recognition tasks have to do with each other.

3. SEGMENTATION VERSUS OBJECT RECOGNITION

In brief, object recognition (or detection) methods can be split into 3 main categories or tenets:

- Learning based approaches
- Segmentation-based approaches
- Local feature-based approaches

Learning based approaches consist in optimizing a decision function over a training set of example data, according to a given error or fidelity criterion. The decision function satisfies a given architecture framework: Artificial Neural and Convolutional Nets (see [3], [4] in the case of faces) and Support Vector Machines. An implicit model of the object or object category is built during the training stage, then used for matching during the recognition stage. Such methods provide impressive results in the case of face detection and their applications to other object having comparable characteristics seems straightforward. This is very attractive on the one hand, since only example and counter examples need to be input in the learning machine, with no prior feature extraction. The counterpart is that one has to develop and apply as many object (category) detectors as object types to recognize. This may lead to a computationally complex indexing stage.

Local feature-based approaches rely on the independent detection of sufficiently discriminative local geometric or color features. This supposes that such distinctive features exist for the object of interest. The main advantages of this approach are: its low computational complexity above all if the descriptors used present some invariance, its robustness with respect to occlusions, and the fact that no prior segmentation is needed. A representative example is given in [5].

Finally, segmentation-based approaches rely on explicit prior segmentation of the image sequence into objects, or at least the extraction of the spatio-temporal area corresponding to the object of interest. This is why it has

been and still is involving much effort, among wich the OSIAM French project [6] and the COST 211 European project [7]. For complex objects, mostly characterized by their structure rather than their surface property, motion appears as the most relevant information to use for segmentation. Statistical models describing the object interior can also be used.

Once the object boundary extracted, labeling is obtained through matching with a reference object model. Such a model can concern the silhouette (or external 2D projected shape). This is the approach underlying the MPEG-7 choice of (invariant) "shape" descriptors such as the Region shape based on the Angular Radial Transform and the Contour Shape based on the Curvature Scale Space [8]. Unfortunately, silhouette is very variable with 3D pose, and is furthermore very sensitive to occlusion. As a matter of fact, in the case of partial occlusion, the external contour is no longer representative of the object to be matched during the retrieval stage.

For most objects or object categories of interest, the inner structure is really the most meaningful. Therefore, another way to exploit the segmentation result is simply to match the segmented object with some model images. In that case, segmentation largely relieves the matching procedure since it solves the issue of scale and rotation compensation, save in the case of severe occlusion and therefore is an alternative to invariant object representation (see [9] for instance). However, this works only if the object at hand has some luminance or color invariance properties, which is the case for faces and in general for natural (biological) entities, but more rarely for human-made (manufactured) objects. In [10] Garrido et Al. have proposed a smart approach to directly exploit lower level hierarchical region-based segmentation of both the candidate images and the query object for object matching and retrieval. Such a region-based segmentation is one possible way to exploit the inner structure of the objects present in the images while relying only on generic image description.

Nevertheless, even in the case of almost perfect object segmentation, the role of prior (unsupervised) segmentation is essentially to limit the search area, scale and orientation, but brings little for the the recognition process itself. In the general case of generic background, camera motion and multiple objects (usual in commercial videos and movies), automatic object segmentation is still an unsolved problem, and semi-automatic approaches [11], [12] will be preferred still for long. In that case, relying on prior segmentation, independent of any recognition process, to solve the problem of position, scale and orientation compensation, is very questionable and may lower the reliability of the whole process. Besides, accurate object boundary location is certainly not necessary.

4. TYPE-STYLE AND FONTS

In this paper, we have attempted to show why indexing of image sequences based on object labeling is a very important stage for a fully automated indexing with a high level of semantic value, that indeed corresponds to end users expectations. Paradoxically, this does not imply that explicit segmentation of the video into the objects that compose it is needed and can help the subsequent recognition or retrieval stage. Except in the favorable case of unoccluded purely 2D objects, the external shape is not sufficient to characterize the entity. Internal structure matching is in some way indispensable and segmentation can help only if it is quite reliable, which is not yet the case in general. We therefore think that segmentation and indexing target 2 separate purposes.

9. REFERENCES

[1] José M. Martínez, “ISO/IEC JTC1/SC29/WG11 N4674: MPEG-7 Overview,” *Approved Document*.

[2] Remco C. Veltkamp, Mirela Tanase, *"Content-Based Image Retrieval Systems: A Survey "*, IEEE Image processing, Vol. 1, N° 1, Oct. 2001, pp 100-148.

[3] Féraud R., Bernier O., Viallet, J.E., Collobert M., "A Fast and Accurate Face Detector for indexation of Faces Images, In Proc. Of FG'2000 Conference on Automatic Face and Gesture Recognition, March 2000, Grenoble, France.

[4] Garcia C., Delakis M., “Training Convolutional Filters for Robust Face Detection”, Proceedings of the IEEE International Workshop of Neural Networks for Signal Processing (NNSP2003), Toulouse, France, Sept. 2003.

[5] Se S., Lowe D.G., Little J., "Global Localization using distinctive features, International Conference on Intelligent Robots and Systems (IROS'2002), 2002, pp 226-231.

[6] The OSIAM project web site, *http://www.telecom.gouv.fr/rnrt/projets/posiam.htm.*

[7] The COST 211 project web site, *http://www.iva.cs.tut.fi/COST211/introduction/whatis.htm*

[8] Text of ISO/IEC 15938-3/FCD Information Technology - Multimedia Content Description Interface - Part 3 Visual. ISO/IEC JTC1/SC29/WG11. March 2001.

[9] F. Ghorbel, "A complete Invariant Description for Grey-level Images by the Harmonic Analysis Approach, Pattern Recognition Letters, Vol. 15, 1994, pp 1043-1051.

[10] Garrido L., "Hierarchical region based processing of images and video sequences: Application to filtering, segmentation and information retrieval", PhD thesis dissertation, UPC, April 2002.

[11] M. Mazière, F. Chassaing, L. Garrido and P. Salembier, "Segmentation and tracking of video objects for a content-based video indexing context, Proc. of ICME 2000, July 2000.

[12] S. Jehan, M. Barlaud, {G. Aubert, "DREAM²S: Deformable Regions driven by an Eulerian Accurate Minimization Method for image and video Segmentation, International Journal of Computer Vision, 2003.