SEGMENTATION-BASED EXTRACTION OF IMPORTANT OBJECTS FROM VIDEO FOR OBJECT-BASED INDEXING

Muhammet Baştan, Uğur Güdükbay, Özgür Ulusoy

Bilkent University
Department of Computer Engineering
Ankara, Turkey

ABSTRACT

We describe a method to automatically extract important video objects for object-based indexing. Most of the existing salient object detection approaches detect visually conspicuous structures in images, while our method aims to find regions that may be important for indexing in a video database system. Our method works on a shot basis. We first segment each frame to obtain homogeneous regions in terms of color and texture. Then, we extract a set of regional and inter-regional color, shape, texture and motion features for all regions, which are classified as being important or not using SVMs trained on a few hundreds of example regions. Finally, each important region is tracked within each shot for trajectory generation and consistency check. Experimental results from news video sequences show that the proposed approach is effective.

Index Terms— video object extraction, segmentation, important object, indexing

1. INTRODUCTION

Support for detailed object-based spatio-temporal queries in a video database system requires the extraction of important objects from the video for indexing; this is impossible to do manually for video databases of realistic size. Therefore, automatic object extraction is a crucial first step in the processing chain in such systems. Specifically, an MPEG-7 compliant video database system may store video data as follows to enable complex queries. Each video is decomposed into shots. Each shot is decomposed into moving regions corresponding to important objects or regions. Each moving region may have several features attached: color, shape, texture, trajectory, annotation, etc. The static background content of a shot can be represented by decomposing it into keyframes which in turn can be decomposed into still regions having low-level and high level features attached. Using such a system we may answer very complex object-based queries on a video collection (e.g., retrieve video segments in which object *X* with such low-level features, having such a trajectory, and appearing in a scene with such properties); however, the real challenge is to extract the *important objects/regions* from the video automatically. Depending on the homogeneity of objects, extracted regions may correspond to semantic objects (e.g., human), however, current state-of-the-art in computer vision is not yet able to detect semantic objects. Therefore, we use the terms *object* and *region* interchangeably.

We use the term *important object* to define any video object that should be stored in the database because users may be interested in performing queries about it, though it is a bit bold to claim that we can characterize such objects completely and detect them with high accuracy. In section 2, we review the literature on saliency detection briefly, and in section 3, we present the characteristics of *important objects* as we define them. We claim that our definition does also encompass saliency as defined in the literature; this is also supported by the experimental results presented in section 4.

2. RELATED WORK AND MOTIVATION

In the literature, salient objects are defined as the visually distinguishable, conspicuous image components that attract our attention at the first glance. These are usually high contrast regions, or regions with significantly different appearance compared to their surroundings. Detection of salient regions is also referred to as *image attention analysis*.

The first remarkable work on saliency is [1]. It combines multiscale image features into a single topographical saliency map. Using this map and a dynamic neural network, the attended image locations are selected in order of decreasing saliency. In [2], a saliency map is generated based on local contrast analysis, then a fuzzy growing method is used to extract attended areas or objects from the saliency map by simulating human perception. In [3], image segmentation is formulated as the identification of single perceptually most salient structure in the image. In [4], the authors try to obtain OOI (Object-of-Interest) segmentation of natural images into background and a salient foreground by region merging

This work is supported by European Union 6th Framework Program under Grant No. FP6-507752 (MUSCLE Network of Excellence Project) and TÜBİTAK under grant No. EEEAG-105E065 (New Information Society Technologies for Turkey).

978-1-4244-2571-6/08/$25.00 ©2008 IEEE

within a selected attention window based on saliency maps and saliency points from the image. In [5], the log spectrum of each image is analyzed to obtain the spectral residual, which is transformed into spatial domain to obtain the saliency map which in turn indicates the positions of proto-objects. In [6], salient object detection is formulated as an image segmentation problem, in which the salient object is separated from the image background. A set of novel features are proposed: multi-scale contrast, center-surround histogram, and color spatial distribution to describe a salient object locally, regionally, and globally. A Conditional Random Field (CRF) is learned using a human labeled set of training images to effectively combine these features for salient object detection.

There has been little work on salient object detection in video, taking into account the valuable motion information. The model proposed in [7] predicts the saliency of a spatio-temporal event based on the information it contains. The joint spatial and temporal conditional probability distributions of spatio-temporal events are modeled and their spatio-temporal saliencies are computed in an integrated way. Motion channels are added to intensity-based saliency maps in [8]. The authors argue that addition of motion information, as they described, did not improve the performance. In [9], spatial and temporal saliency maps are fused to compute a spatio-temporal saliency map. A spatio-temporal saliency framework is described in [10]; it combines spatial feature detection, feature tracking and motion prediction in order to generate a spatio-temporal saliency map to differentiate predictable and unpredictable motions in video.

Motivation. This work is motivated by the need to meet the requirement of a video database system to detect important objects from video for indexing, which is missing in what have been proposed in the literature. To our knowledge, this work is the first to address this issue. Main characteristics of our approach are as follows.

- definition of important objects from the perspective of a video database system
- segmentation-based approach
- use of regional & inter-regional features instead of frequently used pixel-based features
- diverse, simple to compute feature set: color, texture, shape, motion
- designed to work in video and able to detect multiple objects

3. SEGMENTATION-BASED DETECTION OF IMPORTANT OBJECTS

In contrast to pixel-based saliency map approaches, we take a segmentation-based approach to the detection of important objects with the hope of achieving higher accuracy in terms of object boundaries. We first detect the shot boundaries in the video using a global HSV color histogram based approach, which performs satisfactorily. Then, we work on a shot basis; detect and track important objects within each shot to obtain the trajectories as well. The detection process starts with the spatial segmentation of each frame. Then, features are extracted from each region. The features are of mainly two kinds: (1) features extracted from the regions, (2) features measuring how different a region is from its neighbors and from all other regions. Using the features and with SVMs, each region is classified as being important or not.

3.1. Spatial Segmentation of Frames

Spatial segmentation of frames is a key step in our processing chain, since it directly affects the region properties and hence the final detections. We use the JSeg image segmentation algorithm [11] for this purpose, since it is widely used due to its performance, and it is freely available on the web. In JSeg, images are first quantized to several representative color classes in YUV color space. Then, each pixel is replaced by its representative class label. By applying a “good” segmentation criterion to local windows, a “J-image” is produced. Finally, a region growing approach is used to segment the image based on multi-scale J-images.

3.2. Characteristics of Important Objects & Features

For a region, the notion of being important or not is a subjective matter; different people may select different regions from the same content. We can still agree upon a set of characteristics using some heuristics. We now list these along with possible features to represent each region.

1. In videos, objects in camera focus are usually important (e.g., a speaking head, as in Figure 2, top image). Objects in camera focus have higher contrast and sharper edges compared to the background. This can be measured using region variance, entropy, and edge strength on the region boundary.

2. Visually conspicuous regions may be important. This is indicated by how different the region is from its surrounding, from the rest, and hence can be measured by inter-regional contrast on specific features (e.g., color, texture, motion).

3. Moving regions may be important (e.g., walking person, sailing boat, as in Figure 1, Figure 2); hence velocity is an important clue.

4. Too large, too small, too long/thin regions are usually not important. For example, large regions are usually background. This suggests using area, shape properties.

5. Important objects should be consistent; they should appear in most of the frames within a shot (e.g., 10 % of the frames).

Using these characteristics, we compute the following features for each region and obtain a feature vector of length 18. These are easy to compute once the segmentation is available.

1. Regional color, shape, texture and motion features
   - Region color variance (maximum of 3 RGB channels) and entropy (from greyscale image)
   - Average region velocities in X and Y directions computed by optical flow between successive frames
   - Region area & shape properties: ratio of region area to frame area, aspect ratio, ratio of region area to MBR area (compactness)
2. inter-regional features
   - Sum of difference of mean color, variance, entropy, velocity of a region from its neighbors, and from all other regions, weighted by region areas
   - Boundary edge strength

3.3. Classification of the Regions & Tracking

We selected 300+ positive/negative important region examples, computed features, normalized them to zero mean and unit variance, and trained an SVM with polynomial kernel. Using this SVM, we classify each region as important or unimportant. For each important region, the distance to separating hyperplane returned by the SVM is assigned as the importance score. We rank the regions according to this score and select the first N regions. This parameter can be used to tune the detection precision & recall of the system. The number of important regions as detected by SVM can be zero or more, hence our system can say that there is no important region in the frame.

We track each important region throughout the shot for consistency check and also for trajectory information to store in the database. We keep a list of important tracked regions within each shot. In each frame, we try to find a match for each tracked region by first imposing position and shape constraints and then checking color histogram distance between the regions. At the end of processing a shot, if a region appeared less than a threshold (10 % of the frames), it does not qualify as an important region. This threshold can also be used to tune the detection precision & recall.

4. EXPERIMENTAL RESULTS

We tested our system on several news video sequences with length hundreds of frames each. Figure 1 shows example detections of varying quality. If the frames are easy to segment, so that the segmentation quality is satisfactory, the resulting detections are good. In an example opposite case, as shown in the top-right image of Figure 1, the walking person could not be correctly detected due to poor segmentation.

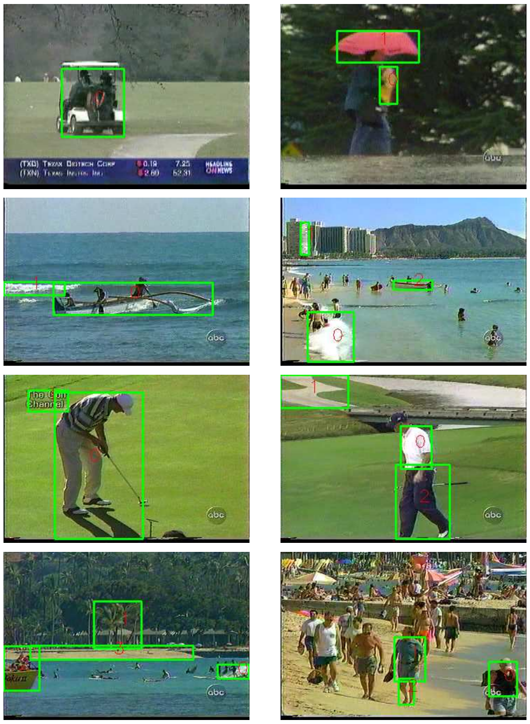

Fig. 1. Example detections. Numbers within rectangles show the rank of saliency for that region.

We compared the performance of our system with one of the leading saliency model (SM) [1] approaches, whose MATLAB implementation is freely available at [12]. Figure 2 shows detection examples by the two methods. We limited N to 5 in the experiments. In most cases, our approach performs much better in terms of human visual perception and in terms of our definition of *important objects*. We also computed the precision-recall values of the two systems on 2 test video sequences with a total of 668 frames. A user is presented the first 5 detected regions which he evaluated as correct/wrong/missed. The evaluation is again based on our definition of *important object*. The precision-recall graph in Figure 3 indicates that our system is significantly better.

5. DISCUSSION AND FUTURE WORK

Experimental results show that the proposed approach is promising in detecting important regions in videos. Obtaining semantically meaningful objects is still a research issue. Using the output of our system and merging the regions by their saliency and motion properties may yield semantically better results. Current set of features are simple, easy to compute, yet have proved to be effective. There is still more work to do for the selection of features and classification methods.

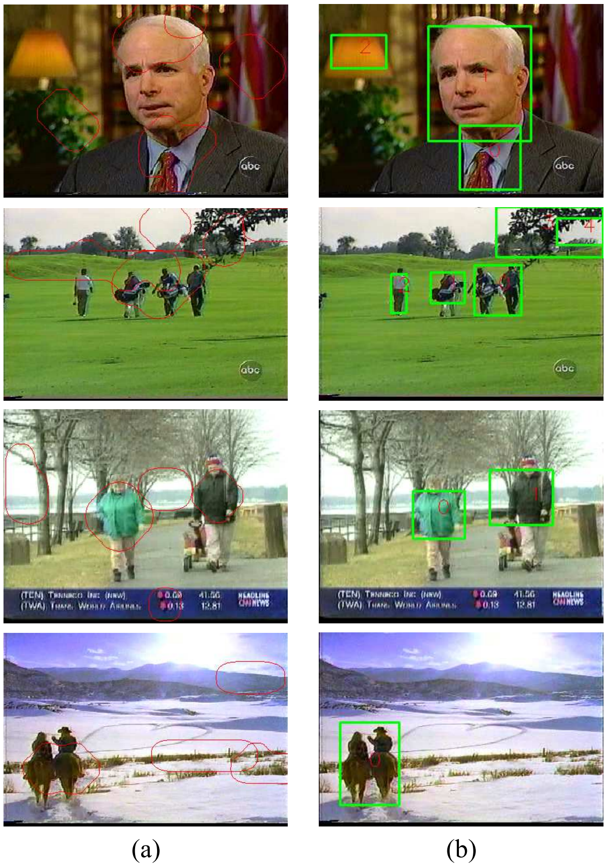

Fig. 2. Visual comparison of first 5 detections. (a) SM, (b) Our approach

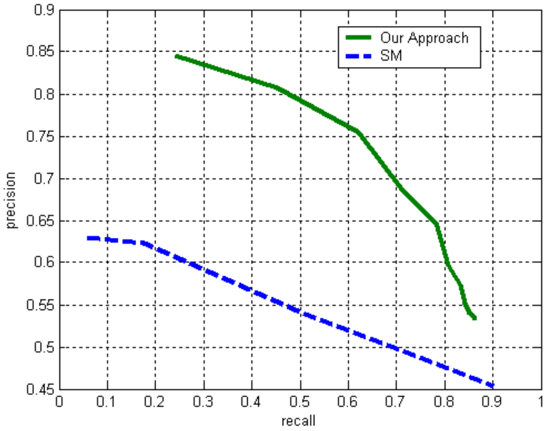

Fig. 3. Precision-recall graph for the detection of first 5 important objects, comparing two approaches.

6. REFERENCES

[1] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[2] Y.F. Ma and H.J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proceedings of the Eleventh ACM International Conference on Multimedia*, 2003, pp. 374–381.

[3] Feng Ge, Song Wang, and Tiecheng Liu, "Image-segmentation evaluation from the perspective of salient object extraction," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, vol. I, pp. 1146–1153.

[4] Byoung Chul Ko and Jae-Yeal Nam, "Automatic object-of-interest segmentation from natural images," in *Proceedings of the 18th International Conference on Pattern Recognition*, 2006, pp. 45–48.

[5] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision Pattern Recognition*, June 2007, pp. 1–8.

[6] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H.Y. Shum, "Learning to detect a salient object," in *IEEE Conference on Computer Vision Pattern Recognition*, June 2007, pp. 1–8.

[7] Guoping Qiu, Xiaodong Gu, Zhibo Chen, Quqing Chen, and Charles Wang, "An information theoretic model of spatiotemporal visual saliency," in *IEEE International Conference on Multimedia and Expo*, July 2007, pp. 1806–1809.

[8] Trent J. Williams and Bruce A. Draper, "An evaluation of motion in artificial selective attention," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2005, vol. 3, p. 85.

[9] O. Le Meur, D. Thoreau, P. Le Callet, and D. Barba, "A spatio-temporal model of the selective human visual attention," in *IEEE International Conference on Image Processing*, September 2005, vol. 3, pp. III–1188–91.

[10] Yang Liu, Christos-Savvas Bouganis, and Peter Y K. Cheung, "A spatiotemporal saliency framework," in *IEEE International Conference on Image Processing*, October 2006, pp. 437–440.

[11] Y. Deng and B.S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800–810, Aug 2001.

[12] Saliency Toolbox, http://www.saliencytoolbox.net.