COMPRESSED-DOMAIN OBJECT DETECTION FOR VIDEO UNDERSTANDING

Vasileios Mezaris[1,2], *Ioannis Kompatsiaris*[2], *and Michael G. Strintzis*[1,2]

[1]Information Processing Laboratory
Electrical and Computer Engineering Dept.
Aristotle University of Thessaloniki
Thessaloniki 54006, Greece

[2]Informatics and Telematics Institute
1st Km Thermi-Panorama Rd,
Thessaloniki 57001, Greece
e-mail: strintzi@eng.auth.gr

ABSTRACT

In this paper, a novel algorithm for the real-time, unsupervised object detection in compressed-domain sequences is proposed. The algorithm utilizes color and motion information present in the compressed stream as well as a simple object model. Extraction of the MPEG-7 dominant color descriptor, clustering of macroblocks to dominant color clusters and model-based cluster selection are employed for object detection in I-frames, while temporal tracking is employed for P-frames. The proposed methodology assumes neither a static camera nor that there exists a single dominant color in the frame, which represents the object of interest. Experimental results of road detection in various sequences demonstrate the efficiency of the proposed approach and reveal the potential of employing it for video understanding and semantic information extraction in context-specific applications.

1. INTRODUCTION

Digital video is an integral part of many newly emerging multimedia applications. New image and video standards, such as MPEG-4 and MPEG-7, do not concentrate only on efficient compression methods but also on providing better ways to represent, integrate and exchange visual information [1]. Although these standards provide the needed functionalities in order to manipulate and transmit objects and metadata, their extraction is out of the scope of the standards and is left to the content developer.

Furthermore, video understanding and semantic information extraction have been identified as an important step towards more efficient manipulation of visual media [2]. To this end, several approaches have been proposed in the literature for segmenting video sequences to objects, which is the first step towards video understanding. These include generic methods operating in both the raw [3] and the compressed domain [4], as well as raw-domain context-specific methods for identifying predefined objects in a video [5].

This work focuses on context-specific object detection in MPEG-2 compressed sequences. It is based on utilizing information found in the compressed stream as well as prior knowledge regarding the object in the form of a simple object model. The latter could be effectively integrated in an *a priori* knowledge representation framework as in [6, 7]. An overview of the proposed scheme is presented in Fig. 1.

The use of the proposed algorithm is demonstrated in the context of road detection. The detection of the road area is an important step towards detecting moving objects of interest in any road sequence (e.g. racing sequences, surveillance using moving cameras) and, most importantly, automatically understanding the semantics of the sequence by efficiently modelling the events captured in it. Applied to sequences of this context, the algorithm is shown to achieve real-time, unsupervised road detection.

The remainder of the paper is organized as follows: in section 2, the extraction of information from the compressed stream is discussed. Section 3 deals with the road model selection. In section 4, the road detection algorithm is developed. Section 5 contains an experimental evaluation of the developed methods, and finally, conclusions are drawn in section 6.

2. COMPRESSED-DOMAIN INFORMATION EXTRACTION

This work is focused on the fast and efficient detection of objects in MPEG-2 compressed streams. The information used by the proposed algorithm is extracted from MPEG sequences during the decoding process. Specifically, the extracted color information is restricted to the DC coefficients of the macroblocks of I-frames, corresponding to the Y, Cb and Cr components of the MPEG color space. These are employed for extracting dominant color regions and selecting the one corresponding to the object of interest, as discussed in the sequel. Additionally, motion vectors are extracted for the P-frames and are used for temporal tracking of the object detected in the I-frames, in the absence of relevant color information. Since P-frames are coded using mo-

This work was supported by the European IST project SCHEMA.

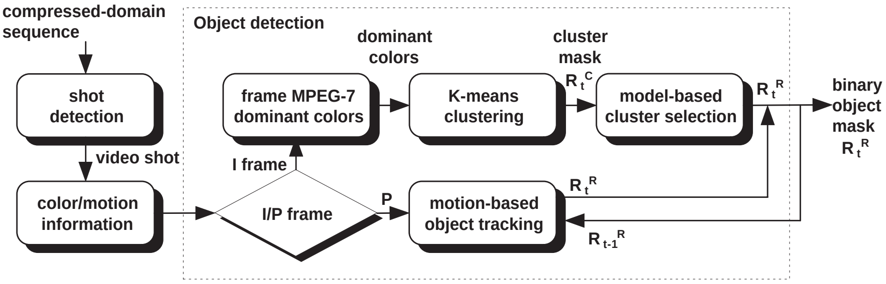

Fig. 1. Compressed-domain object detection algorithm overview.

tion information from I-frame to P-frame or from P-frame to P-frame, their motion information provides a clearer indication of the motion of the region of interest in comparison to motion information derived from temporally adjacent frames. Consequently, the motion information contained in B-frames is disregarded.

3. ROAD MODEL SELECTION

The road area detection problem in a known context (i.e. knowing that the examined sequence is of such content, e.g. car racing, that a road area exists) can be modelled as a dominant color detection problem, as is similarly done in [5] for soccer field detection. However, in [5] it is assumed that there exists a single dominant color in the frame; this dominant color represents the region of interest. This assumption does not necessarily hold for sequences where road (or other object) detection could be of importance, as shown in Fig. 2.

Given that there exist more than one dominant colors in the frame, and assuming that one of them (not necessarily the most dominant one) represents the region of interest (i.e. the road), it is imperative that additional *a priori* knowledge is employed for selecting the appropriate dominant color region. This is accomplished in this work by means of a simple road model. The latter accounts for only color information, since shape is hardly characteristic of the different possible views of the road area, as illustrated in Fig. 2 and in the experimental results section. Thus, the road area is modelled as three independent normal distributions, each corresponding to one component of the color space. The parameters (μ_k, σ_k), $k \in \{1, 2, 3\}$ of the model are estimated by averaging the corresponding values calculated for the members of a training set, i.e. a number of manually generated road areas belonging to various sequences. To account for variability in the color of the different road areas under different conditions (e.g. lighting), this model is not used for directly detecting road macroblocks; instead, it is used for selecting one of a number of dominant color clusters, as discussed in the next section.

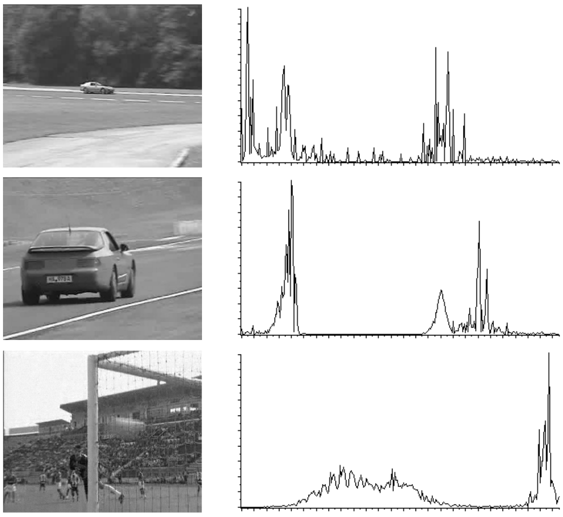

Fig. 2. Histograms for two frames of a car racing sequence and one frame of a soccer sequence. In all cases, there exist more than one dominant colors.

4. ROAD AREA DETECTION ALGORITHM

The proposed algorithm for road detection is based on exploiting the color and motion information of the MPEG-2 macroblocks and consists of two main procedures:

- Road detection in I-frames, using DC color information and prior knowledge in the form of the aforementioned simple road model.
- Road detection in P-frames, using macroblock motion vectors and the output of road detection in preceding I frames.

These procedures are discussed in more detail in the sequel.

4.1. I-frame processing

The procedure for detecting the road area in I-frames consists of three main steps:

- Step 1. Extraction, using the DC coefficients of the I-frame, of the MPEG-7 Dominant Color Descriptor.
- Step 2. Clustering of the frame macroblocks to dominant color clusters.
- Step 3. Model-based selection of the cluster corresponding to the road area.

Initially (step 1), the DC coefficients of the I-frame F_t are used for extracting the Dominant Color Descriptor [8]. This is a standardized descriptor, part of the MPEG-7 Visual Standard [1], which compactly describes the representative colors of the frame. The descriptor is defined as:

$$DCD = \{\{c_i, p_i, v_i\}, s^h\}, \ \ i = 1, \ldots, N, \ \ N \leq 8$$

where c_i is the ith dominant color, p_i is a percentage expressing its degree of dominance, v_i is its color variance (optional field) and s^h is a value indicating the spatial homogeneity of the dominant colors.

According to our assumption, the road area is represented by one of these dominant colors. However, no assumption has been made involving the degree of dominance of the particular color, e.g. it is not assumed that the N_H most dominant colors have a higher probability of representing the road area than the N_L less dominant ones. Thus, the p_i values can be disregarded and all N dominant colors are used for initializing a simple K-means algorithm, similarly to [9]. Applied to the frame macroblocks (step 2), this produces a macroblock-level accuracy mask R_t^C containing N dominant color clusters. The similarity of each cluster with the road model is then to be evaluated.

Model-based selection of the cluster corresponding to the road region (step 3) is performed using the Earth Mover's Distance (EMD) [10]. The EMD computes the distance between two distributions, which are represented by signatures, and is defined as the minimum amount of work needed to change one signature into the other. The notion of "work" is based on the user-defined ground distance, which is the distance between two features; in this work, the Euclidean distance is employed to this end.

The signatures involved in the computation of the EMD are defined as:

$$S = \{s_j = (\mathbf{m}_j, w_j)\}$$

where $\mathbf{m}_j$ represents a d-dimensional point (e.g. the three mean color values corresponding to a histogram bin) and w_j is the weight associated with this point (e.g. the non-zero value of the corresponding histogram bin; empty bins can be omitted). For each dominant color cluster, its histogram is calculated and is treated as its signature. Regarding the road model signature, a set of a few points in the three-dimensional color space and the corresponding non-zero values of the continuous model $\{(\mu_k, \sigma_k)\}$ are easily extracted, given the continuous model. The cluster for which the model-cluster EMD is minimum is selected as representative of the road area.

4.2. P-frame processing

In order to detect the road area in P-frames, in the absence of color information, temporal tracking of macroblocks is performed using the motion information associated with them in the compressed stream and the road mask extracted for the preceding frame. The temporal tracking is based upon the work presented in [11], where objects were manually marked by selecting their constituent macroblocks and these objects were subsequently tracked in the compressed domain using the macroblock motion vectors.

Let R_{t-1}^R denote the road mask derived from frame F_{t-1} and let $\tau(.)$ be the tracking operator defined in [11], taking as input a macroblock at time $t-1$ and outputting the corresponding macroblock or macroblocks at time t. Then, the operator $\mathcal{T}(.)$ can be defined as taking a road mask R_{t-1}^R as input, applying the $\tau(.)$ operator to all macroblocks of that mask, and outputting the road mask at time t, as estimated by temporal tracking. Thus,

$$R_t^R = \mathcal{T}(R_{t-1}^R), \ \ F_t \in \mathcal{P}$$

where $\mathcal{P}$ is the set of P-frames.

5. EXPERIMENTAL RESULTS

The proposed method was tested on a variety of video sequences. A number of frames and corresponding road segmentation masks R_t^R, which show the road in white and the non-road macroblocks in black, are presented in Fig. 3. It can be seen from these that the algorithm has succeeded in detecting the real road area depicted in the sequences. Very few macroblocks have been falsely classified.

Additionally, the proposed approach succeeds in adding minimal computational overhead to the computational complexity of a standard MPEG decoder. In particular, the compressed domain road detection algorithm presented (excluding any processes of the MPEG decoder and the storage of the road masks, which is algorithm-independent and unnecessary in many cases, e.g. for further analysis and the subsequent inference of semantics) requires on average 60 msec per processed I/P-frame on an 800Mhz Pentium III. This translates to approximately 50 frames per second considering the presence of two consecutive B-frames between every two I/P-frames, which is typical for MPEG-2 sequences and is the case for the employed test media.

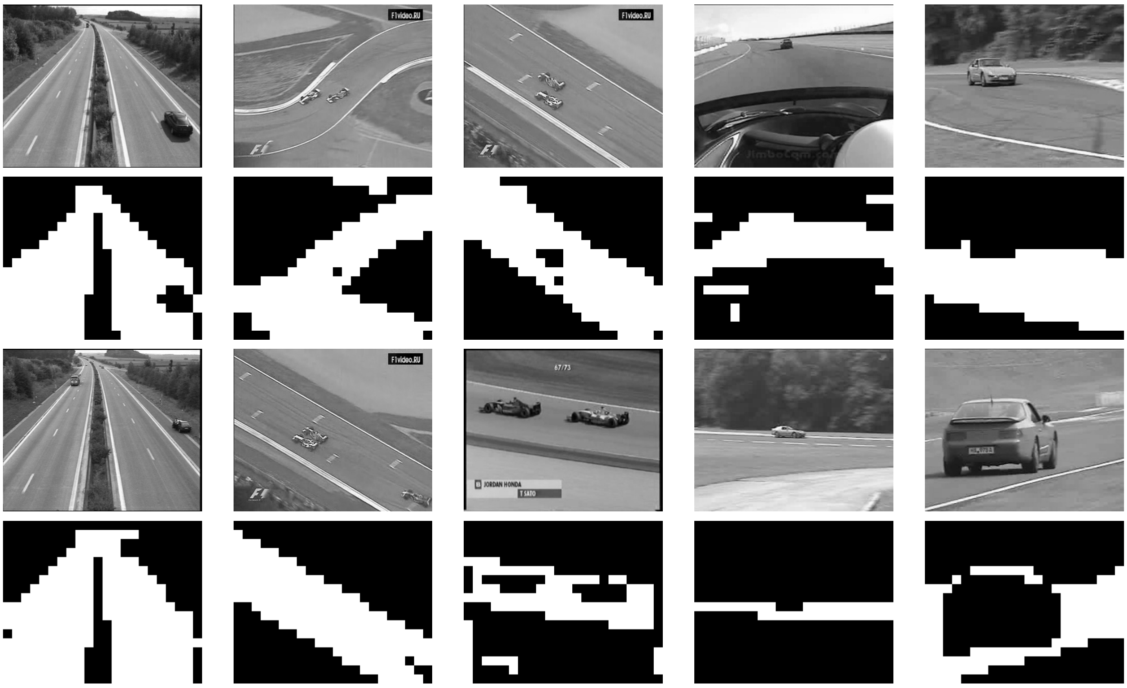

Fig. 3. Results of road detection for various sequences. Road macroblocks are shown in white.

6. CONCLUSIONS

An algorithm for the context-specific, unsupervised object detection in MPEG-2 sequences was presented in this paper. The algorithm was employed for road detection and was shown to perform in real-time on a PC, producing semantically meaningful results, which can be further employed for video understanding and semantic knowledge extraction. Due to its real-time, unsupervised operation, the proposed algorithm is appropriate for context-specific applications requiring the manipulation of large volumes of visual data.

7. REFERENCES

[1] T. Sikora. The MPEG-7 Visual standard for content description - an overview. *IEEE Trans. on Circuits and Systems for Video Technology, special issue on MPEG-7*, 11(6):696–702, June 2001.

[2] S.-F. Chang. The holy grail of content-based media analysis. *IEEE Multimedia*, 9(2):6–10, Apr.-Jun. 2002.

[3] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. Video Object Segmentation using Bayes-based Temporal Tracking and Trajectory-based Region Merging. *IEEE Trans. on Circuits and Systems for Video Technology, to appear*.

[4] V. Mezaris, I. Kompatsiaris, E. Kokkinou, and M.G. Strintzis. Real-time compressed-domain spatiotemporal video segmentation. In *Proc. Third International Workshop on Content-Based Multimedia Indexing (CBMI03)*, Rennes, France, Sept. 2003.

[5] A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Trans. on Image Processing*, 12(7):796–807, July 2003.

[6] A. Yoshitaka, S. Kishida, M. Hirakawa, and T. Ichikawa. Knowledge-assisted content based retrieval for multimedia databases. *IEEE Multimedia*, 1(4):12–21, Winter 1994.

[7] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. An Ontology Approach to Object-based Image Retrieval. In *Proc. IEEE Int. Conf. on Image Processing (ICIP03)*, Barcelona, Spain, Sept. 2003.

[8] B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. on Circuits and Systems for Video Technology, special issue on MPEG-7*, 11(6):703–715, June 2001.

[9] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. A framework for the efficient segmentation of large-format color images. In *Proc. International Conference on Image Processing*, volume 1, pages 761–764, 2002.

[10] Y. Rubner, C. Tomasi, and L.J. Guibas. A Metric for Distributions with Applications to Image Databases. In *Proc. IEEE International Conference on Computer Vision*, pages 59–66, Bombay, India, Jan. 1998.

[11] L. Favalli, A. Mecocci, and F. Moschetti. Object tracking for retrieval applications in MPEG-2. *IEEE Trans. on Circuits and Systems for Video Technology*, 10(3):427–432, Apr. 2000.