A KNOWLEDGE-BASED APPROACH TO DOMAIN-SPECIFIC COMPRESSED VIDEO ANALYSIS

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 54124, 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 approach to domain-specific video analysis is proposed. The proposed approach is based on exploiting domain-specific knowledge in the form of an ontology to detect video objects corresponding to the semantic concepts defined in the ontology. The association between the visual objects and the defined semantic concepts is performed by taking into account both qualitative attributes of the semantic objects (e.g. color homogeneity), indicating necessary preprocessing methods (color clustering, respectively), and numerical data generated via training (e.g. color models, also defined in the ontology). To enable fast and efficient processing, this methodology is applied to MPEG-2 video, requiring only its partial decoding. The proposed approach is demonstrated in the domain of Formula-1 racing video and shows promising results.

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 has been identified as an important step towards more efficient manipulation of visual media [2]. In well-structured specific domain applications (e.g. sports and news broadcasting) domain specific features that facilitate the modelling of higher level semantics can be extracted [3, 4]. A priori knowledge representation models are used as a knowledge base that assists semantic-based classification and clustering [5, 6]. In [7], semantic entities, in the context of the MPEG-7 standard, are used for knowledge-assisted video analysis and object detection, thus allowing for semantic level indexing. In [8], fuzzy ontological relations and context aware fuzzy hierarchical clustering are employed to interpret multimedia content for the purpose of automatic thematic categorization of multimedia documents. In [9] the problem of bridging the gap between low-level representation and high-level semantics is formulated as a probabilistic pattern recognition problem.

This work focuses on context-specific detection of objects in MPEG-2 compressed sequences. It is based on utilizing information found in the compressed stream as well as prior knowledge regarding the objects in the form of an ontology. The proposed approach is demonstrated in the context of Formula-1 racing video. The detection of semantically significant objects, such as the road area and the cars in such racing video is an important step towards understanding the semantics of a temporal segment of the video by efficiently modelling the events captured in it. Applied to sequences of this context, the proposed approach shows promising results in exploiting the supplied domain-knowledge for achieving fast and unsupervised detection of objects.

The remainder of the paper is organized as follows: in section 2, the use of knowledge for domain-specific analysis is discussed, along with the building of a domain-specific ontology suitable for analysis. Section 3 deals with compressed video processing and its integration with the knowledge in the employed ontology. Section 4 contains an experimental evaluation of the developed methods, and finally, conclusions are drawn in section 5.

2. KNOWLEDGE FOR DOMAIN-SPECIFIC ANALYSIS

As opposed to generic video, where various objects may be depicted in it and the detection of any one of them may or may not be important for eventually extracting a semantic interpretation for a given sequence of frames, in domain-specific video there tends to be a small number of known objects that to a great extent can reveal the semantics of the sequence. This reveals the potential of employing a priori knowledge for detecting this limited number of objects.

In this work, this a-priori knowledge is modelled as an ontology, which defines the expected semantic concepts in terms of qualitative attributes, numerical data generated via training and relations among them. This makes possible the definition of object attributes as a function of the corresponding attributes of another object (e.g. for object "car", its size can be defined as $size_{car} < a \cdot size_{road}$).

A simplified ontology for the Formula-1 domain is presented in Fig. 1. This defines three objects of interest for the domain-specific analysis: the road, the grass typically found on the side of the road, and the racing car. A size relation between the road and car concepts is additionally defined. The differences in the object definitions (e.g. homogeneity, connectivity) indicate the dif-

This work was supported by the EU projects aceMedia "Integrating knowledge, semantics and content for user centred intelligent media services" (FP6-001765) and SCHEMA "Network of Excellence in Content-Based Semantic Scene Analysis and Information Retrieval" (IST-2001-32795).

ferent processing methods that should be used for finding possible matches in a frame (e.g. color clustering only or additional motion-based cluster merging, the application or not of a four-connectivity component labelling algorithm, etc.). Further, these heterogeneous definitions clearly demonstrate the generality of the proposed approach (e.g. the same analysis can be applied for semantic information extraction from a football game by assigning the appropriate properties to concepts such as players, ball, field, etc.).

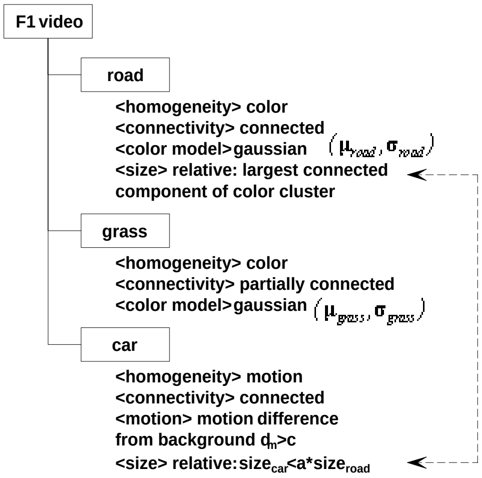
F1 video
road
<homogeneity> color
<connectivity> connected
<color model>gaussian $(\mu_{road}, \sigma_{road})$
<size> relative: largest connected component of color cluster
grass
<homogeneity> color
<connectivity> partially connected
<color model>gaussian $(\mu_{grass}, \sigma_{grass})$
car
<homogeneity> motion
<connectivity> connected
<motion> motion difference from background d_m>c
<size> relative: $size_{car}$<a*$size_{road}$

Fig. 1. Formula-1 racing video ontology. The ontology can be expanded so as to include additional semantic objects as well as additional knowledge for the existing ones.

In the ontology model used, color 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 identified regions belonging to various sequences. To account for color variability 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 will be explained in more detail in Section 3.2, the Earth Mover's Distance (EMD) will be used for matching the color models stored in the ontology with the video objects extracted using the analysis algorithms defined for each concept.

This approach differs from techniques where a significant region is first extracted and then analysis is based on this result. The road area detection problem, for example, could be modelled in a known context as a dominant color detection problem, as is similarly done in [10] for soccer field detection. However, in [10] it is assumed that there exists a single dominant color in the frame and this dominant color represents the region of interest. This assumption does not necessarily hold for other sequences where for example in racing, road (or also other object, such as grass) detection could be of importance, as shown in Fig. 2.

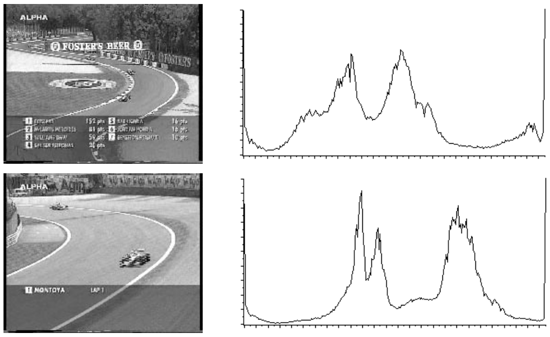

Fig. 2. Histograms for two frames of an F1 racing sequence. In both cases, there exist more than one dominant colors.

3. COMPRESSED VIDEO PROCESSING

3.1. Compressed-domain Information Extraction

To enable the efficient processing of large volumes of visual data, the proposed approach is applied to 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 color clustering, as discussed in the sequel. Additionally, motion vectors are extracted for the P-frames and are used for generating motion information for the I-frames via interpolation. P-frame motion vectors are also necessary for realizing the temporal tracking in P-frames, of the objects detected in the I-frames, as in [11]. Due to the limited information that is required by the proposed approach, only partial decompression of the video stream is necessary.

3.2. I-frame processing

The procedure for detecting the desired objects in I-frames, for which both color information and motion information can be extracted, as previously discussed, starts by performing a set of initial clusterings, to be used in the sequel. Specifically for color, clustering is performed by identifying up to eight dominant colors in the frame, as done by the MPEG-7 Dominant Color descriptor [12], and using them to initialize a simple K-means algorithm, similarly to [13]. The resulting preprocessing mask R_t^{NC} contains a number of non-connected color-homogeneous regions that can be used for model-based selection of "non-connected" semantic objects for which the "homogeneity" property of the ontology is based on color. Applying a four connectivity component labelling algorithm to it, preprocessing mask R_t^{CC} is generated. This is used for model-based selection of "connected" semantic objects for which the "homogeneity" property of the ontology is again based on color. Partial connectivity requirements (e.g. "color-homogeneous object *b* may be represented by more than one connected component, but each such should account for at least $\beta\%$ of the total area of the color cluster") or other requirements can be enforced by combining the information in masks R_t^{NC} and R_t^{CC} to generate suitable preprocessing masks in accordance with the descriptions (e.g. "partially connected") used in the ontology.

Once the sequence is analyzed and appropriate preprocessing masks are produced, a suitable matching method with the models stored in the ontology must be defined. Color-model-based selection of the area corresponding to a color-homogeneous semantic object is performed using a suitable preprocessing mask and the Earth Mover's Distance (EMD) [14]. 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 examined area of the appropriate preprocessing mask (e.g. R_t^{NC}, R_t^{CC}), its histogram is calculated and is treated as its signature. Regarding the color-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 area for which the model-cluster EMD is minimum is selected as representative of the semantic object and is marked with a distinct label in final mask R_t^F.

For motion-homogeneous objects, a similar process is followed, starting with the generation of a preprocessing mask R_t^{MH} containing motion-homogenous areas. Although a motion based clustering procedure could be employed to this end, this is simply done in this work by means of motion based grouping of the connected components of mask R_t^{CC}. Subsequently, the model-based selection depends on the information contained in the ontology; for the "car" object in our experiments, this was restricted to the minimum required motion difference from the background. Assuming a moving camera, the motion difference calculation was preceded by global motion estimation from the macroblock motion vectors, using the bilinear motion model and an iterative rejection procedure [15].

3.3. P-frame processing

In order to detect semantic objects in P-frames, in the absence of color information, temporal macroblock tracking can be performed using the motion information associated with them in the compressed stream and the final mask R_{t-1}^F extracted for the preceding frame. In this work, the temporal tracking is based upon the work presented in [16], 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.

4. EXPERIMENTAL RESULTS

The proposed method was tested on a variety of Formula-1 racing sequences of 720×576 pixels. A number of frames and corresponding final segmentation masks R_t^F, showing the grass in light gray, the road in dark gray and the most likely car area in black, are presented in Fig. 3. In these masks, macroblocks identified as belonging to neither one of these three classes are shown in white. It can be seen from these results that the algorithm has succeeded in employing the knowledge contained in the ontology for detecting the real objects depicted in the sequences.

Additionally, the proposed approach succeeds in adding minimal computational overhead to the computational complexity of a standard MPEG decoder. In particular, the compressed domain video processing (excluding any processes of the MPEG decoder and the storage of the segmentation masks, which is algorithm-independent and unnecessary in many cases, e.g. for further analysis and the subsequent inference of semantics) requires on average 0.62 sec per processed I-frame on a Pentium IV PC.

5. CONCLUSIONS

An approach for the context-specific, unsupervised object detection in MPEG-2 sequences was presented in this paper. The algorithms were applied to Formula-1 racing video and were shown to produce semantically meaningful results, which can be further employed for video understanding and knowledge extraction. Due to its time-efficient, unsupervised operation, the proposed algorithm is appropriate for context-specific applications requiring the manipulation of large volumes of visual data. This knowledge-driven approach can be easily adapted to different domains by appropriately defining the necessary concepts and their properties.

Future work includes the selection of a formal language for representing the ontology (e.g. RDF Schema), additional and more complex model representations of features and semantic reasoning based on the ontology.

6. 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*, vol. 11, no. 6, pp. 696–702, June 2001.

[2] S.-F. Chang, "The holy grail of content-based media analysis," *IEEE Multimedia*, vol. 9, no. 2, pp. 6–10, Apr.-Jun. 2002.

[3] A. Yoshitaka and T. Ichikawa, "A survey on content-based retrieval for multimedia databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 1, pp. 81–93, Jan/Feb 1999.

[4] W. Al-Khatib, Y.F. Day, A. Ghafoor, and P.B. Berra, "Semantic modeling and knowledge representation in multimedia databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 1, pp. 64–80, Jan/Feb 1999.

[5] A. Yoshitaka, S. Kishida, M. Hirakawa, and T. Ichikawa, "Knowledge-assisted content based retrieval for multimedia databases," *IEEE Multimedia*, vol. 1, no. 4, pp. 12–21, Winter 1994.

[6] 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, vol. II, pp. 511–514.

[7] G. Tsechpenakis, G. Akrivas, G. Andreou, G. Stamou, and S.D. Kollias, "Knowledge-Assisted Video Analysis and Object Detection," in *Proc. European Symposium on Intelligent Technologies, Hybrid Systems and their implementation*

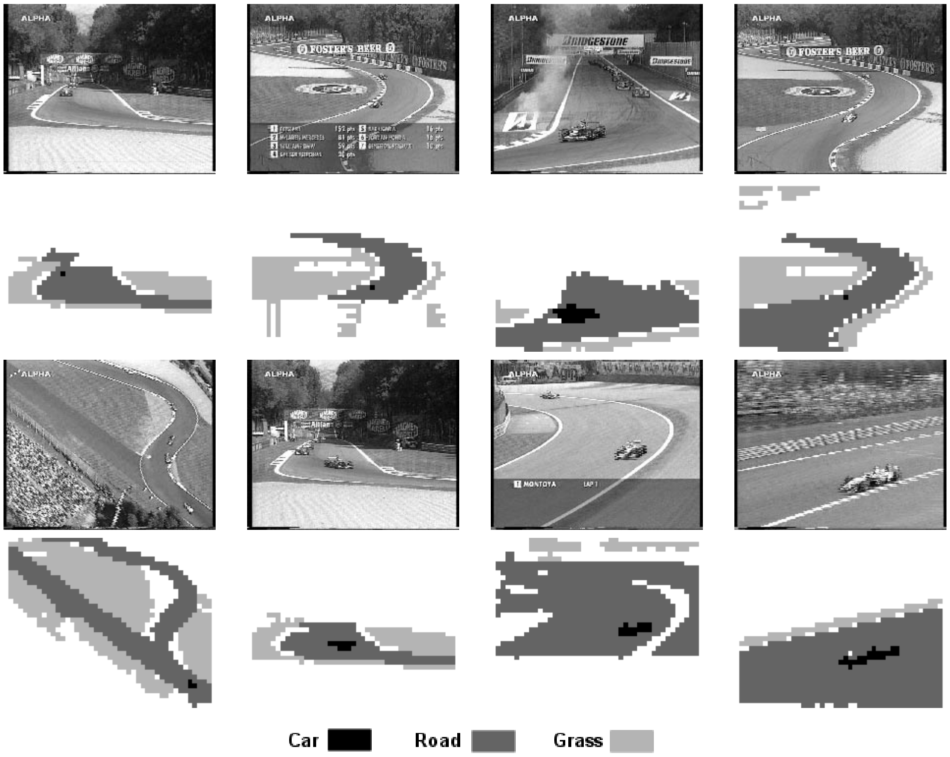

Fig. 3. Results of grass (light gray), road (dark gray), and single car (black) detection for F1 video. Macroblocks identified as belonging to neither one of the three classes are shown in white.

on Smart Adaptive Systems (Eunite02), Algarve, Portugal, September 2002.

[8] M. Wallace, G. Akrivas, P. Mylonas, Y. Avrithis, and S. Kollias, "Using Context and Fuzzy Relations to Interpret Multimedia Content," in *Proc. 3rd Int. Workshop on Content-Based Multimedia Indexing, CBMI03*, Rennes, France, September 22-24 2003.

[9] M. Ramesh Naphade, I.V. Kozintsev, and T.S. Huang, "A factor graph framework for semantic video indexing," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 1, pp. 40–52, Jan. 2002.

[10] A. Ekin, A.M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. on Image Processing*, vol. 12, no. 7, pp. 796–807, July 2003.

[11] 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, pp. 373–380.

[12] 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*, vol. 11, no. 6, pp. 703–715, June 2001.

[13] 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*, 2002, vol. 1, pp. 761–764.

[14] 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*, Bombay, India, Jan. 1998, pp. 59–66.

[15] T. Yu and Y. Zhang, "Retrieval of video clips using global motion information," *Electronics Letters*, vol. 37, no. 14, pp. 893–895, July 2001.

[16] L. Favalli, A. Mecocci, and F. Moschetti, "Object tracking for retrieval applications in MPEG-2," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 3, pp. 427–432, Apr. 2000.