Knowledge-Assisted Video Analysis for Content-Adaptive Coding and Transmission

Vasileios Mezaris[1], Nikolaos V. Boulgouris[2], Ioannis Kompatsiaris[1]

[1]Informatics and Telematics Institute
Centre for Research and Technology Hellas
1st Km Thermi-Panorama Road
Thessaloniki 57001, Greece

[2]Department of Electronic Engineering
Division of Engineering
King's College London
London WC2R 2LS, U.K.

Abstract

Efficient video content management and exploitation, including coding and transmission, can greatly benefit from taking into account the semantics of the video content. However, the extraction of the latter is a non-trivial task that involves associating low-level features of the image domain with high-level semantic descriptions. In this paper, a knowledge-assisted approach for extracting semantics of domain-specific video content is presented and is employed for effecting content-adaptive coding and transmission of the video. To this end, domain knowledge considers both low-level features (color, motion, shape) and spatial behavior (topological and directional information) of video content for the purpose of analysis, as well as domain-, application- and user-specific importance factors associated with each domain concept that guide coding and transmission on the basis of the output of the analysis. Examples of the application of the proposed approach to sports videos, specifically belonging to the Formula One and Tennis domains, are provided in the Results section.

1 Introduction

Recent advances in computing technologies have made available vast amounts of digital video content, leading to ever-increasing flow of audiovisual information. This results in a growing demand for efficient video content management and exploitation, including coding and transmission. A key enabling factor for this is the acquisition of higher-level information about the meaning of the video content, which however is a non-trivial problem.

The existing difficulty [1] in mapping semantic concepts as perceived by humans into a set of automatically extracted low-level image features, can be alleviated to some extent for a particular application domain by means of domain specific knowledge. Different approaches have been used for the implementation of particular parts of the domain knowledge such as formal knowledge representation theories, semantic web technologies, Dynamic Belief networks etc. For example, in [4], semantic web technologies are used, while in [16] *a priori* knowledge representation models are used as a knowledge base that assists semantic-based classification and clustering. In [7], an object ontology coupled with a relevance feedback mechanism is introduced, in [14], semantic entities, in the context of the MPEG-7 standard, are used for knowledge-assisted video analysis and object detection, while in [10], associating low-level representations and high-level semantics is formulated as a probabilistic pattern recognition problem.

Once a semantic interpretation of the video is extracted, this can be used for effecting content-adaptive video coding and transmission. Specifically, we present a novel methodology for the coding and transmission of video. Our method allows the optimization of video delivery based on the video content and the significance of its constituent objects.

In this paper a knowledge-assisted, domain-specific video analysis framework is introduced and employed for content-adaptive video coding and transmission. The analysis framework uses a genetic algorithm to support efficient object localization and identification. An initial segmentation generates automatically a set of atom-regions and subsequently their low-level descriptors are extracted. Analysis is then performed using the necessary processing tools and by relating high-level symbolic representations included in the ontology to visual features extracted from the signal domain. Additionally, the genetic algorithm decides how the atom-regions should be merged in order to form objects in compliance with the object models defined in the domain ontology. The output of this analysis process is then used for the optimization of the coding and transmission of the

First International Workshop on Semantic Media Adaptation and Personalization (SMAP'06)
0-7695-2692-6/06 $20.00 © 2006 IEEE

video.

The remainder of the paper is structured as follows: section 2 considers domain ontology development and section 3 contains a presentation of the segmentation and descriptor extraction algorithms and discusses the implementation of the genetic algorithm. Section 4 addresses the issues of coding and transmission exploiting the previously generated analysis results. Indicative results are presented in section 5 and finally, conclusions are drawn in section 6.

2 Domain Knowledge

The knowledge about the examined domain is encoded in the form of an ontology. The developed ontology includes the objects that need to be detected, their visual features and their spatiotemporal relations, as well as domain-, application- and user-specific importance factors associated with each domain concept. These descriptions provide the system with the required knowledge to find the optimal interpretation for each of the examined video scenes, i.e. the optimal set of mappings among the available atom-regions and the corresponding domain-specific semantic definitions, and to subsequently employ them for guiding efficient coding and transmission. To account for objects of no interest that may be present in a particular domain and for atom-regions that fail to comply with any of the object models included in the ontology, the unknown object concept is introduced; this concept is assigned the m minimum of the domain-, application- and user-specific importance factors. In addition, support is provided for the definition of associations between low-level descriptions and the algorithms to be applied for their extraction. In the following, a brief description of the main classes is presented.

Class **Object** is the superclass of all objects to be detected during the analysis process: when the ontology is enriched with the domain specific information it is subclassed to the corresponding domain salient objects. Class **Object Interrelation Description** describes the objects spatiotemporal behavior, while **Low-Level Description** refers to the set of their representative low-level visual features. Since real-world objects tend to have multiple different instantiations, it follows that each object prototype instance can be associated with more than one spatial description and respectively multiple low-level representations. Different classes have been defined to account for the different types of low-level information (color, shape, motion etc.). These are further subclassed to reflect the different ways to represent such a feature (e.g. color information could be represented by any of the color descriptors standardized by MPEG-7, the distribution models of the respective color space etc.) The actual values that comprise the low-level descriptors (e.g. the DC value elements, color space) are under the **Low-Level Descriptor Parameter** class.

Class **Importance Factors** is the main class containing knowledge about the coding of the given domain object. It is subclassed to classes **Domain-specific Importance Factor**, **Application-specific Importance Factor** and **User-specific Importance Factor**, which define the importance factor values I_d, I_a and I_u. During coding, these are combined using an appropriate function $f(I_d, I_a, I_u)$ to drive the coding process.

Providing domain-specific spatiotemporal information proves to be particularly useful for the identification of specific objects, since it allows discrimination of objects with similar low-level characteristics as well as of objects whose low-level features alone are not adequate for their identification. The applied spatial relations consider two-dimensional, binary relations, defined between regions with connected boundaries. In the current implementation the included spatial relations are the eight topological relations resulting from the 9-intersection model as described in earlier works on spatial relations representation and reasoning [12, 3], enhanced by the four relative directional relations, i.e. right, left, above, below. The used low-level descriptors are the MPEG-7 Dominant Color and Region Shape descriptors, the motion norm of the averaged global motion-compensated block motion vectors for each region blocks and the ratio between a region's area and the square of its perimeter (compactness).

3 Knowledge-Assisted Video Analysis

3.1 Color and motion initial segmentation

The color segmentation is based on the extraction of up to eight dominant colors in the frame, as proposed in the MPEG-7 Dominant Color descriptor [6], used to initialize a simple K-means algorithm as detailed in [8].

The motion segmentation is based on a two step algorithm. The first step follows the segmentation methodology of [7], considering a block matching approach, in order to obtain a coarse but very fast segmentation. Indeed, an iterative rejection scheme [17] based on the bilinear motion model is used to effect foreground/background segmentation. Meaningful foreground spatiotemporal objects are formed by initially examining the temporal consistency of the output of iterative rejection, clustering the resulting foreground macroblocks to connected regions and finally performing region tracking. Furthermore, this first step provides an efficient estimation of the 8 parameters of the bilinear camera motion model. As a second step, the previous motion segmentation is used to initialize a region-based motion segmentation algorithm based on smoothing spline active contours [11]. Smoothing splines offer a robust active contour implementation to overcome the problem of noisy data that working with MPEG streams implies. Hence, improved accuracy over the first step motion segmentation is achieved. Furthermore, the contour defining the extracted moving regions is given by a parametric equation which allows a fast computation for geometric curve features such as perimeter, area, or moments, involved in the low-level feature descriptor extraction.

The generated color and motion segmentation masks are merged giving priority to color information. That is to say, if a motion-based segmented region consists of two or more color-based segmented atom-regions, this region is split according to the color segmentation. Finally, a region-based smoothing spline active contour is applied to the resulted segmentation mask in order to provide the parametric contour equation of each atom-region.

3.2 Low-level descriptors extraction

The low-level descriptors defined in section 2 are extracted for each atom-region as follows. We compute the Dominant Color descriptor applying the MPEG-7 eXperimentation Model (XM) [6]. The region motion feature, based on the aforementioned motion segmentation algorithm, is defined by the norm of the average global-motion-compensated motion vectors evaluated on the blocks belonging to the atom-region considered. To extract the compactness descriptor, we compute the area and the perimeter of each region using a fast algorithm, proposed in [5], based on spline properties of the parametric contour description.

3.3 Genetic Algorithm

As previously mentioned, the initially applied color and motion segmentation algorithms, result in a set of over-segmented atom-regions. Assuming for a single image N_R atom regions and a domain ontology of N_O objects, there are $N_R^{N_O}$ possible scene interpretations. To overcome the computational time constraints of testing all possible configurations, a genetic algorithm is used [9]. Genetic algorithms (GAs) have been widely applied in many fields involving optimization problems, as they proved to outperform other traditional methods. They build on the principles of evolution via natural selection: an initial population of individuals (chromosomes encoding the possible solutions) is created and by iterative application of the genetic operators (selection, crossover, mutation) the optimal, according to the defined fitness function, solution is reached.

In our framework, each individual represents a possible interpretation of the examined scene, i.e. the labelling of all atom-regions either as one of the considered domain objects or as unknown. An object instantiation is identified by its corresponding concept and an identifier used to differentiate instances of the same concept. The domain ontology contains information about the maximum allowed number of detected instances for each object. In order to reduce the search space, the initial population is generated by allowing each gene to associate the corresponding atom-region only with those objects that the particular atom-region is most likely to represent. For example in the domain of Tennis a green atom-region may be interpreted as a Field, Wall or Unknown object but not as Ball or Player. Therefore, for each individual included in the initial population, the corresponding gene is associated with one of the three aforementioned object concepts (instead of the available N_O). The set of plausible candidates for each atom-region is estimated according to the low-level descriptions included in the domain ontology.

The following functions are defined to estimate the degree of matching in terms of low-level visual and spatial features respectively between an atom-region r_i and an object concept o_j.

- the interpretation function $\mathcal{I}_M^t(r_i, o_j)$, assuming that gene g_t associates region r_i with object o_j, to provide an estimation of the degree of matching between o_j and r_i. $\mathcal{I}_M^t(r_i, o_j)$ is calculated using the descriptor distance functions realized in the MPEG-7 XM and is subsequently normalized so that $\mathcal{I}_M^t(r_i, o_j)$ belongs to $[0, 1]$, with a value of 1 indicating a perfect match.

- the interpretation function $\mathcal{I}_R^t(r_i, o_j, r_k, o_l)$, which provides an estimation of the degree to which the spatial relation between atom-regions r_i and r_k satisfies the relation $\mathcal{R}$ defined in the ontology between objects o_j, o_l to which r_i and r_k are respectively mapped to by gene g_t.

Since each individual represents the scene interpretation, the Fitness function has to consider the above defined low-level visual and spatial matching estimations for all atom-regions. As a consequence the employed Fitness function is defined as follows:

$$Fitness(g_t) = (\sum_{i}^{N_R} \mathcal{I}_M^t(r_i, o_m)) \prod_{i}^{N_R} \prod_{j \in S_i} \mathcal{I}_R^t(r_i, o_m, r_j, o_l)$$

where S_i denotes the set of neighboring atom-regions of r_i, since the spatial relations used have been defined only for regions with connected boundaries as mentioned in 2. It follows from the above definitions that the optimal solution is the one that maximizes the Fitness function. This process elegantly handles the merging of atom-regions: any neighboring such regions belonging to the same object according to the generated optimal solution are simply merged. In our implementation, the following genetic operators were used: roulette wheel selection, in which individuals are given a probability of being selected that is directly proportionate to their fitness and uniform crossover, where genes of the parent chromosomes are randomly copied.

4 Content-Adaptive Coding and Transmission

The availability of a systematic way for the semantic description of video sequences provides new means to deal with the subsequent delivery of video.

We assume that each object is compressed using a embedded coding method. This means that the object is represented using a scalable stream that can be decoded at arbitrary source rates depending on the required quality. However, in a practical video transmission scenario, the decoding quality does not depend only on source coding but also on channel coding. Therefore, for the ith object we denote the decoded quality as $D_i(\rho_i, n)$, where ρ_i is the source+channel bitrate that is devoted to the source and channel coding of the ith object, and n denotes the bit error rate of the Binary Symmetric Channel over which the video sequence is transmitted.

Based on the above, we define the total distortion function as

$$D = \sum_{i=0}^{N-1} f_i D_i(\rho_i, n) \tag{1}$$

where f_i is the relative importance of the $i$the object, as defined in section 2. The total bitrate for the coding and transmission of the video sequence is

$$R = \sum_{i=0}^{N-1} \rho_i \tag{2}$$

Since $\rho_i = s_i + c_i$, where s_i and c_i are the source and channel bits for the coding of the ith object, the total rate can be expressed as

$$R = \sum_{i=0}^{N-1} (s_i + c_i) \tag{3}$$

Using the above formulation, the allocation of bits to the objects of the ontology can be achieved using Lagrangian methods. Object-wise optimization of rate allocation is possible if the optimal source and channel rates for each block are known. In practice, this can be achieved using the techniques in [13], i.e., by solution of an unconstrained problem which aims to the minimization of an objective function F of the form

$$F = D + \lambda_L R \tag{4}$$

where D and R and are given by (1) and (3), respectively, and λ_L is a Lagrange multiplier. A similar optimization for the blockwise coding of images was presented in [2].

5 Experimental Results

The proposed approach was tested on a variety of Formula One and Tennis domain videos. As illustrated in Fig.1 the system output is a segmentation mask outlining the semantic interpretation, i.e. a mask where different colors representing the objects defined in the ontology are assigned to each of the produced regions. The objects of interest included in each domain ontology are illustrated in table 1. For all experimental domains, the low-level descriptors values included in the corresponding knowledge base were extracted from a training set of manually annotated images.

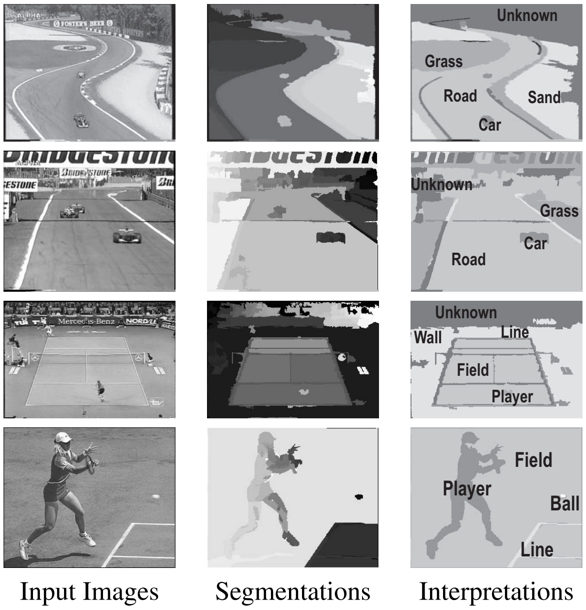

Figure 1. Formula One and Tennis domain analysis results

The time required for performing the previously described tests was between 5 and 10 seconds per frame, excluding the process of motion information extraction via block matching for which efficient and inexpensive hardware implementations exist [15]. More specifically, the time to perform pixel-level segmentation was about 2 seconds, while the time required by the genetic algorithm to reach an optimal solution varied depending on the number of atom-regions and the number of spatial relations. The extraction of the low-level and spatial descriptions is performed before the application of the genetic algorithm. In general, the proposed approach proved to produce satisfactory results as long as the initial color-based segmentation did not segment two objects as one atom-region.

The subsequent coding and transmission, using the techniques described previously, appeared to benefit from the novel content-aware approach. Specifically, objects designated as being of higher importance were consistently decoded at higher qualities than those of the other objects in the ontology.

Table 1. Formula One and Tennis domain objects of interest

Domain	Concept
Formula One domain	Road
	Car
	Sand
	Grass
Tennis domain	Field
	Player
	Line
	Ball
	Wall

6 Conclusions

In this paper, a knowledge-assisted domain-specific video analysis approach, which exploits the fuzzy inference capabilities of a genetic algorithm, is employed for supporting content-adaptive video coding and transmission. Domain knowledge includes both low-level features and spatial relations of video content for the purpose of analysis, as well as domain-, application- and user-specific importance factors associated with each domain concept to guide coding and transmission. The developed domain ontology provides a flexible conceptualization that allows the easy addition of new concepts, low-level and spatiotemporal descriptors, as well as updated importance factors, thus supporting different abstraction levels and flexible adaptation of the analysis and coding process to different domains, applications and users.

Acknowledgment

This work was supported by the EU project aceMedia (FP6-001765).

References

[1] W. Al-Khatib, Y. Day, A. Ghafoor, and P. Berra. Semantic modeling and knowledge representation in multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):64–80, Jan/Feb 1999.

[2] N. Boulgouris, N. Thomos, and M. Strintzis. Transmission of Images Over Noisy Channels Using Error-Resilient Wavelet Coding and Forward Error Correction. *IEEE TCSVT*, Dec 2003.

[3] M. J. Egenhofer and R. D. Franzosa. Point set topological relations. *International Journal of Geographical Information Systems*, 5:161–174, 1991.

[4] J. Hunter, J. Drennan, and S. Little. Realizing the hydrogen economy through semantic web technologies. *IEEE Intelligent Systems Journal - Special Issue on eScience*, 19:40–47, 2004.

[5] M. Jacob, T. Blu, and M. Unser. An exact method for computing the area moments of wavelet and spline curves. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):633–642, 2001.

[6] B. Manjunath, J.-R. Ohm, 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.

[7] V. Mezaris, I. Kompatsiaris, N. Boulgouris, and M. Strintzis. Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(5):606–621, May 2004.

[8] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Still Image Segmentation Tools for Object-based Multimedia Applications. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(4):701–725, June 2004.

[9] M. Mitchell. *An introduction to Genetic Algorithms*. MIT Press., 1996.

[10] M. R. Naphade, I. Kozintsev, and T. Huang. A factor graph framework for semantic video indexing. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(1):40–52, Jan. 2002.

[11] F. Precioso, M. Barlaud, T. Blu, and M. Unser. Smoothing b-spline active contour for fast and robust image and video segmentation. In *ICIP (1)*, pages 137–140, 2003.

[12] D. A. Randell, Z. Cui, and A. G. Cohn. A spatial logic based on regions and connection. In *KR*, pages 165–176, 1992.

[13] Y. Shoham and A. Gersho. Efficient Bit Allocation for an Arbitrary Set of Quantizers. *IEEE Trans. on Acoust., Speech, Signal Processing*, 36:1445–1453, Sep. 1988.

[14] G. Tsechpenakis, G. Akrivas, G. Andreou, G. Stamou, and S. Kollias. Knowledge-Assisted Video Analysis and Object Detection. In *Proc. European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (Eunite02)*, Algarve, Portugal, September 2002.

[15] J.-C. Tuan, T.-S. Chang, and C.-W. Jen. On the data reuse and memory bandwidth analysis for full-search block-matching VLSI architecture. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(1):61–72, January 2002.

[16] 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.

[17] T. Yu and Y. Zhang. Retrieval of video clips using global motion information. *Electronics Letters*, 37(14):893–895, July 2001.