A Genetic Algorithm-Based Approach to Knowledge-Assisted Video Analysis

N. Voisine[2], S. Dasiopoulou[1,2], F. Precioso[2], V. Mezaris[1,2], I. Kompatsiaris[2], M. G. Strintzis[1,2]

[1]Information Processing Laboratory
Electrical and Computer Engineering Dept.
Aristotle University of Thessaloniki
540 06 Thessaloniki, Greece

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

Abstract— **Efficient video content management and exploitation requires extraction of the underlying semantics, a non-trivial task associating low-level features of the image domain and high-level semantic descriptions. In this paper, a knowledge-assisted approach for extracting semantics of domain-specific video content is presented. Domain knowledge considers both low-level features (color, motion, shape) and spatial behavior (topological and directional information). During the preprocessing step, a set of over-segmented homogenous atom-regions is generated and their low-level and spatial descriptions are extracted. A genetic algorithm is then applied in order to find the optimal interpretation according to a specific domain conceptualization. The proposed approach was tested on the Formula One, Tennis and beach vacations domains showing promising results.**

I. 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 segmentation/analysis methods for extracting semantic information from such content, since the acquisition of higher-level information in terms of meaning is the key enabling factor for the management and exploitation of video content. However, due to the possible different interpretations and intended uses, the ambiguity that is inherent in visual information renders the development of faster hardware or the evolution of classic segmentation algorithms insufficient. This difficulty [1] in mapping semantic concepts as perceived by humans into a set of automatically extracted low-level image features, can be alleviated 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 [2], semantic web technologies are used, while in [3] *a priori* knowledge representation models are used as a knowledge base that assists semantic-based classification and clustering. In [4], an object ontology coupled with a relevance feedback mechanism is introduced, in [5], semantic entities, in the context of the MPEG-7 standard, are used for knowledge-assisted video analysis and object detection, while in [6], associating low-level representations and high-level semantics is formulated as a probabilistic pattern recognition problem.

In this paper a knowledge-assisted, domain-specific video analysis framework is presented, using a genetic algorithm to support efficient object localization and identification. The localization and identification of the domain salient objects are prerequisites for extracting accurate semantic information. An initial segmentation generates automatically a set of atom-regions and subsequently their low-level descriptors are extracted. Analysis may then be performed by 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. Following this approach, the detection of the important objects depends largely on the knowledge base of the system and as a result it can be easily applied to different domains provided that the knowledge base is enriched with the respective domain knowledge.

The remainder of the paper is structured as follows: section II considers domain ontology development, section III contains a presentation of the segmentation and descriptor extraction algorithms, while in section IV the implementation of the genetic algorithm is discussed. Experimental results are presented in section V and finally, conclusions are drawn in section VI.

II. 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. 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. 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. 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.

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 [7], [8], enhanced by the four relative directional relations, i.e. right, left, above, below. Other qualitative spatial relations such as near, far, between etc. can be easily defined combining the ones previously mentioned. Symmetricity and transitivity properties as well as the inverse of each of the defined spatial relations are specified. Consequently, more complex spatial descriptions can be inferred, reducing at the same time the number of required explicit descriptions. 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).

Enriching the ontology with domain specific knowledge results in populating the ontology with appropriate instances, i.e. prototypes for the objects to be detected. The presented system interprets the provided information (i.e. the low-level and spatial relation descriptions) as a conjunctive normal form clause, with one conjunct for each description category. Each conjunct is the disjunction of the respective category descriptors associated with the particular prototype instance. As will be explained in section IV, fuzzy matching criteria are incorporated in the fitness function used to determine the plausibility of each interpretation. The followed approach proves twofold advantageous as it also allows to tackle the inevitable loss of objects connectivity in the 2D image plane: over-segmented atom-regions belonging to the same object class are appropriately merged to form a single instance of the respective concept.

III. Preprocessing

A. 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 [9], used to initialize a simple K-means algorithm as detailed in [10].

The motion segmentation is based on a two step algorithm. The first step follows the segmentation methodology of [4], considering a block matching approach, in order to obtain a coarse but very fast segmentation. Indeed, an iterative rejection scheme [11] 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 [12]. 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.

B. Low-level descriptors extraction

The low-level descriptors defined in section II are extracted for each atom-region as follows. We compute the Dominant Color descriptor applying the MPEG-7 eXperimentation Model (XM) [9]. 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 [13], based on spline properties of the parametric contour description.

Another spline contour property provides the means to cope with one of the main issues of the region shape descriptor extraction. Indeed, the Angular Radial Transform involved in the MPEG-7 region shape computation considers the evaluation for each region of specific normalized central moments. These moments are defined considering that each region is included into the unit disk. This normalization process can be efficiently managed using the spline structure affine invariance, since a spline curve subjected to an affine transformation is still a spline curve whose parameters (the control points) are obtained by subjecting the original spline curve control points to that affine transformation. Thus the first geometric moments of the considered region (its area, its centroid,...) can be computed using fast algorithms [13] in order to evaluate the parameters of the affine transformation corresponding to the normalized contour and then extract the MPEG-7 region shape descriptor.

IV. 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 [14]. 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 II. 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.

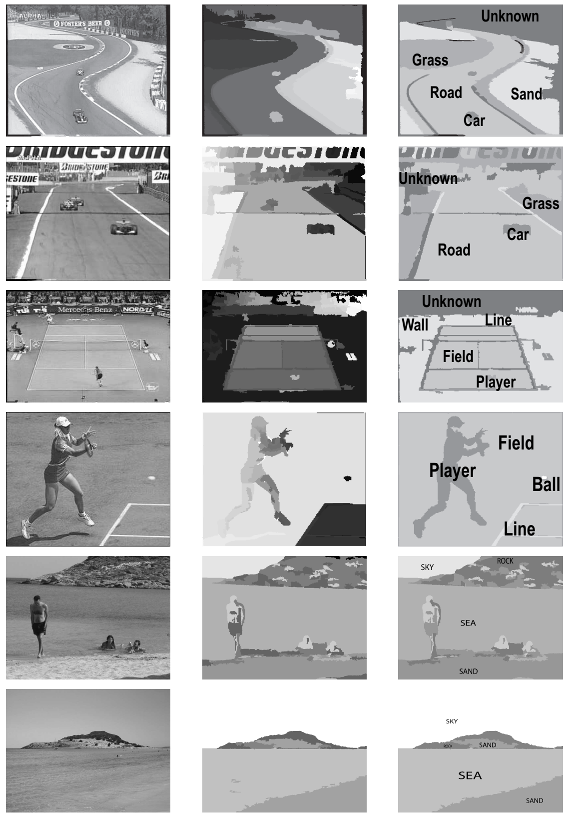
V. Experimental Results

The proposed approach was tested on a variety of Formula One and Tennis domain MPEG-2 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 along with their low-level models and spatial relations are illustrated in table I. 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.

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.

TABLE I
FORMULA ONE, TENNIS AND BEACH VACATIONS DOMAIN DEFINITIONS
(dominant color descriptor (DC), motion descriptor (MOV), compactness descriptor (CPS), externally connected relation (EC), below relation (BEW) and inclusion relation (INC))

Concept	Visual models	Spatial relations
Road	$DC^1_{road} \vee DC^2_{road} \vee DC^3_{road}$	Road EC Grass,Sand
Car	$MOV^1_{car} \wedge CPS^1_{car}$	Car INC Road
Sand	$DC^1_{sand} \vee DC^2_{sand}$	Sand EC Grass, Road
Grass	$DC^1_{grass} \vee DC^2_{grass} \vee DC^3_{grass}$	Grass EC Road,Sand
Field	$DC^1_{field} \vee DC^2_{field} \vee DC^3_{field}$	Field EC Wall
Player	MOV^1_{Player}	Player INC Field
Line	$DC^1_{line} \wedge CPS^1_{line}$	Line INC Field
Ball	$DC^1_{Ball} \wedge CPS^1_{Ball}$	Ball INC Field
Wall	$DC^1_{Wall} \vee DC^2_{Wall} \vee DC^2_{Wall}$	Wall EC Field
Sea	$DC^1_{sea} \vee DC^2_{sea} \vee DC^3_{sea}$	Sea BEW Sky
Sand	$DC^1_{sand} \vee DC^2_{sand}$	Sand EC Sea, BEW Sky
Sky	$DC^1_{sky} \vee DC^2_{sky} \vee DC^3_{sky}$	Sky $B\bar{E}W$ Sea, Sand



Input Images Segmentations Interpretations

Fig. 1. Formula One, Tennis and Beach vacations domain results

VI. CONCLUSIONS

In this paper, a knowledge-assisted domain-specific video analysis approach, which exploits the fuzzy inference capabilities of a genetic algorithm, is presented. Domain knowledge includes both low-level visual descriptors and spatial interrelations, and is encoded in the form of an ontology. The genetic algorithm provides a fundamentally different framework compared to knowledge-based systems using formal rules. By encoding the object models defined in the ontology in the form of constraints (fitness function definition), a global optimal interpretation of the examined scene is reached. The developed domain ontology provides a flexible conceptualization that allows the easy addition of new low-level and spatiotemporal descriptors, i.e. supports different abstraction levels.

ACKNOWLEDGMENT

This work was supported by the EU projects SCHEMA (IST-2001-32795) and aceMedia (FP6-001765).

REFERENCES

[1] 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*, 11(1):64–80, Jan/Feb 1999.

[2] 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.

[3] 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.

[4] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. 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.

[5] 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 on Smart Adaptive Systems (Eunite02)*, Algarve, Portugal, September 2002.

[6] 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*, 12(1):40–52, Jan. 2002.

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

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

[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.