KNOWLEDGE-ASSISTED VIDEO ANALYSIS USING A GENETIC ALGORITHM

N. Voisine[2]*, S. Dasiopoulou*[1,2]*, V. Mezaris*[1,2]*,E. Spyrou*[3]*, T. Athanasiadis*[3]*,*
I. Kompatsiaris[2]*,Y. Avrithis*[3]*, M. G. Strintzis*[1,2]

[1]Information Processing Laboratory, Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Greece
[2]Informatics and Telematics Institute/Centre for Research and Technology Hellas, Thessaloniki, Greece
[3]Image, Video and Multimedia Systems Laboratory, School of Electrical and Computer Engineering,
National Technical University of Athens, Greece

ABSTRACT

Efficient video content management and exploitation requires extraction of the underlying semantics, which is a non-trivial task involving the association of low-level features with high-level concepts. In this paper, a knowledge-assisted approach for extracting semantic information of domain-specific video content is presented. Domain knowledge considers both low-level visual features (color, motion, shape) and spatial information (topological and directional relations). An initial segmentation algorithm generates a set of over-segmented atom-regions and a neural network is used to estimate the similarity distance between the extracted atom-region descriptors and the ones of the object models included in the domain ontology. A genetic algorithm is applied then in order to find the optimal interpretation according to the domain conceptualization. The proposed approach was tested on the Tennis and Formula One domains with promising results.

1. INTRODUCTION

Recent advances in computing technologies have made available vast amount of digital video content resulting in a growing research interest in extracting semantic information from such content in order to enable efficient management and exploitation. However, due to the possible different interpretations and intended uses of video resources, the inherent ambiguity in visual information renders the development of faster hardware or the evolution of classic segmentation algorithms insufficient. The difficulty [1], in mapping concepts as perceived by humans (e.g. objects, events) into a set of automatically extracted image features can be alleviated for a particular application domain by means of domain knowledge. Among the different approaches that have been used for implementing particular parts of the domain-specific knowledge are formal knowledge representation theories, semantic web technologies, dynamic belief networks etc. In [2], for example, semantic web technologies are used for representing domain knowledge, while in [3] internal knowledge representation models have been developed. An object ontology coupled with a relevance feedback mechanism is introduced in [4], while in [5] semantic entities in the context of the MPEG-7 standard are defined for knowledge-assisted video analysis and object detection. Finally, in [6], the association of 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 that uses a genetic algorithm to support efficient object localization and recognition is presented. An initial segmentation generates a set of over-segmented atom-regions and subsequently their low-level descriptors are extracted. Based on these descriptors and the ones of the object prototype instances included in the domain ontology, a distance measure is estimated using a neural network that considers all employed descriptors with different weight on each. In the following, the genetic algorithm is applied in order to decide how the initially generated atom-regions should be merged and labelled in order to form meaningful objects in compliance with the ones defined in the domain ontology. Analysis may then be performed by using the necessary processing tools and by relating high-level symbolic representations of the domain ontology to visual features extracted from the signal domain. Following this approach, the detection of the important objects depends largely on the knowledge base of the system and consequently 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 2 considers domain ontology development, section 3 contains a presentation of the applied segmentation and descriptor extraction algorithms, while in section 4 the implementation of the genetic algorithm is discussed. The intelligent distance estimation based on low-level descriptors is presented in section 5. Experimental results are presented in section 6 and finally, conclusions are drawn in section 7.

2. DOMAIN KNOWLEDGE

The knowledge about the examined domain has been encoded in the form of an ontology. The developed ontology includes the objects that need to be detected, their low-level visual features and their corresponding spatial relations. Thus, the corresponding prototype instances provide the system with the knowledge required 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. The domain ontology contains also information about the maximum allowed number of detected instances for each object. In addition, support is provided for defining associations between the defined low-level visual and spatial descriptors and the algo-

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

rithms 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, this class is subclassed to the corresponding domain salient objects. Class **Object Interrelation Description** models the possible object spatiotemporal relations, 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 (temporal) description and respectively multiple low-level representations. The different types of visual information, i.e. color, motion etc,. comprise different classes, which are further subclassed to reflect the different ways to calculate a visual feature (e.g. the color descriptor could be 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 etc. related to the MPEG-7 dominant color descriptor) are under the **Low-Level Descriptor Parameter** class.

In the current implementation the supported spatial relations are: adjacency, inclusion and the four relative directional relations (right, left, above, below), built on Allen's interval algebra [7]. The used visual low-level descriptors are the MPEG-7 dominant color descriptor, the motion norm of the averaged global motion-compensated block motion vectors and compactness defined as the ratio between a region's area and the square of its perimeter. For convenience, the following abbreviations are used for the rest of the paper to refer to the above mentioned low-level and spatial descriptions: dominant color descriptor **(**DC**)**, motion descriptor **(**MOV**)**, compactness descriptors **(**CPS**)**, adjacency relation **(**ADJ**)**, below relation **(**BEW**)** and inclusion relation **(**INC**)**.

Enriching the ontology with domain specific knowledge results in populating the system knowledge base with prototype instances of the objects to be detected. The proposed system interprets the provided information, i.e. the low-level visual features and the spatial relations, as a conjunctive normal form clause consisting of two clauses, one for each description category. Furthermore, each conjunct is defined as the disjunction of the object prototype descriptors belonging to the respective category. To tackle the inevitable loss of objects connectivity in the 2D image plane, atom-regions belonging to the same object are treated as a single instance of the respective concept as long as they satisfy appropriate topological conventions.

3. INITIAL SEGMENTATION AND LOW-LEVEL DESCRIPTORS EXTRACTION

Under the proposed framework, a set of over-segmented atom-regions is generated by combining the color and motion segmentation masks of the preprocessing step. Color segmentation is realized by identifying up to eight dominant colors in the frame, as done by the MPEG-7 dominant color descriptor [8], and using them to initialize a simple K-means algorithm, as in [9]. Motion segmentation is based on extracting motion information for the image sequence [10], and then applying to this motion information the segmentation methodology of [4]. If a motion-based segmented region consists of two or more color-based segmented atom-regions, then it is accordingly split.

The low-level descriptors defined in section 2 are extracted for each atom-region as follows. For the extraction of the dominant color descriptor, the MPEG-7 eXperimentation Model (XM) is employed [8]. Motion information calculation is based on the aforementioned block motion vector estimation using block matching and the calculation of the norm of the averaged global-motion-compensated motion vectors for the blocks belonging to the region. Global motion compensation is based on estimating the 8 parameters of the bilinear motion model for camera motion, using an iterative rejection procedure [11]. To extract the compactness descriptor, the area and the perimeter of the region are calculated.

4. GENETIC ALGORITHM

As previously mentioned, the initially applied color and motion segmentation algorithms result in a set of over-segmented atom-regions. Assuming 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 [12]. Genetic algorithms (GAs) have been widely applied in many fields involving optimization problems, as they have proved to outperform other traditional methods. GAs are built 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) an optimal solution is reached, according to the defined fitness function.

In our framework, each individual represents a possible interpretation of the examined scene, i.e. the labels for the generated atom-regions. An object instantiation is identified by the concept label and an identifier used to differentiate instances of the same concept. 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 correspond to one of the Field, Wall or Unknown Object concepts but not to the Ball or Player ones. The set of valid 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 similarity distance between a region and an object model in terms of their low-level visual and spatial features respectively:

- the interpretation function $\mathcal{I}_M(g_i) \equiv \mathcal{I}_M(R_i, om_j)$. Assuming that g_i associates region r_i with object o_j having model om_j, $\mathcal{I}_M(g_i)$ provides an estimation of the degree of matching between om_j and r_i. $\mathcal{I}_M(R_i, om_j)$ is calculated using the descriptor distance functions realized in the MPEG-7 XM and is subsequently normalized so that $\mathcal{I}_M(R_i, om_j)$ belongs to $[0, 1]$.
- the interpretation function $\mathcal{I}_\mathcal{R}$, which provides an estimation of the degree to which a relation $\mathcal{R}$ holds between two atom-regions.

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 applied Fitness function is defined as follows:

$$Fitness(G) = \sum_{g_i} \mathcal{I}_M(g_i) + \sum_k \sum_{(g_i,g_j),g_i \mathcal{R}_k g_j} \mathcal{I}_{\mathcal{R}_k}(g_i, g_j)$$

where $\mathcal{I}_M(g_i)$ is the estimation function of gene g_i regarding low-level visual similarity and $\mathcal{I}_{\mathcal{R}_k}(g_i, g_j)$ is the estimation function of spatial similarity between g_i and g_j in terms of $\mathcal{R}_k$. 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. 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 concept of unknown object is introduced.

5. INTELLIGENT LOW-LEVEL DESCRIPTORS DISTANCE ESTIMATION

The implementation of the interpretation function $\mathcal{I}_M$ used for the fitness function is explained in more details in this section. Matching of an atom region with an object model is based on the estimation of the distance between the associated low-level descriptors presented in section 2. When the task is to compare two regions based on a single descriptor, several distance functions can be used. In this approach however, the comparison should consider all three low-level descriptors proposed in section 2, with different weight on each. The problem is not trivial because there is not a unique way to compute this distance.

The proposed way to achieve this is based on a back-propagation neural network with a single hidden layer. The network's input consists of the low-level descriptions of both of an atom-region and an object model, while its output is the normalized distance between the atom-region and the model, based on all available descriptors. A training set is constructed using the descriptors of a set of manually labelled atom-regions and the descriptors of the corresponding object models. The network is trained under the assumption that the distance of an atom-region that belongs to the training set is minimum for the associated object and maximum for all others.

When the unknown atom-regions are presented to the trained network along with the description of the objects, the network responds with an estimation of their distance. This distance is then used for the interpretation function $\mathcal{I}_M$, which is used in the fitness function proposed in section 4. An example of average distances between atom-regions and object models is depicted in 1. Although in this case the network is trained only on 145 atom-regions of two frames of a Formula One video sequence and tested on 65 regions of another frame of a different sequence, it is evident that it can generalize and provide a robust estimator of a complex distance function. This is important, especially as manual labelling of the training set is not an easy task.

6. EXPERIMENTAL RESULTS

The proposed approach was tested on a variety of Formula One and Tennis domain MPEG-2 videos. As illustrated in figures 1 and 2, 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 2. In both domains, the low-level descriptors values included in the corresponding knowledge base were extracted from a training set of manually annotated images.

Atom-region	Distance to object model			
	Car	Grass	Road	Sand
Car	0.30	0.71	0.71	0.93
Grass	0.91	0.55	0.75	0.65
Road	0.93	0.96	0.54	0.77
Sand	0.79	0.99	0.73	0.21

Table 1. Distances between atom-regions and object models estimated by the neural network

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 [10]. 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. Additionally, the use of spatial relations proved beneficial, especially for objects whose low-level visual descriptors are quite similar.

Concept	Visual models	Spatial relations
Road	$DC^1_{road} \vee DC^2_{road} \vee DC^3_{road}$	Road ADJ Grass,Sand
Car	$MOV^1_{car} \wedge CPS^1_{car}$	Car INC Road
Sand	$DC^1_{sand} \vee DC^2_{sand}$	Sand ADJ Grass, Road
Grass	$DC^1_{grass} \vee DC^2_{grass} \vee DC^3_{grass}$	Grass ADJ Road,Sand
Field	$DC^1_{field} \vee DC^2_{field} \vee DC^3_{field}$	Field ADJ 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 ADJ Field

Table 2. Formula One and Tennis domain definitions

7. 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, and the adaptation to different domains.

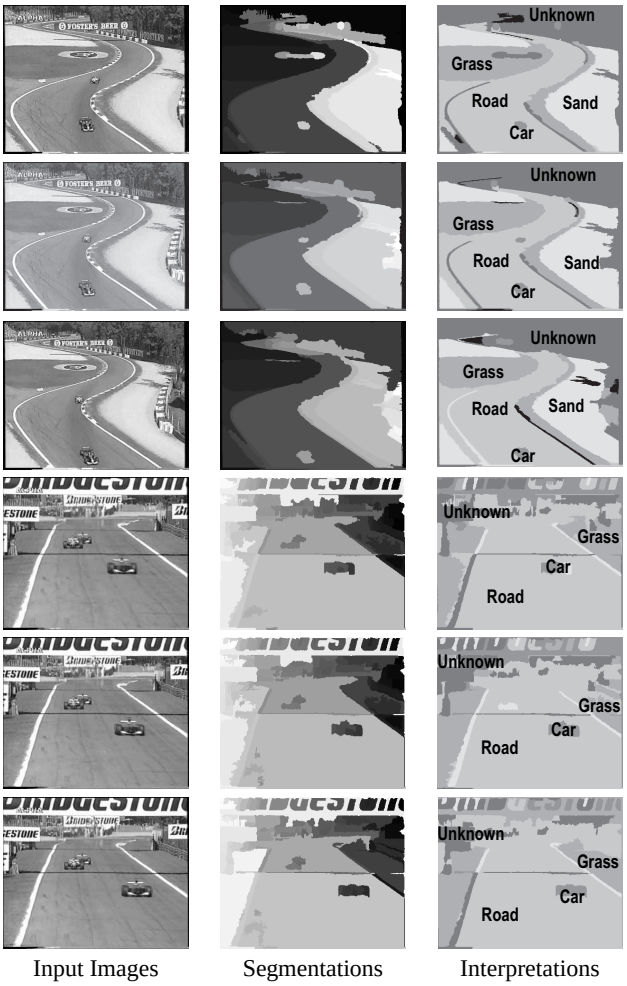

Fig. 1. Formula One domain results

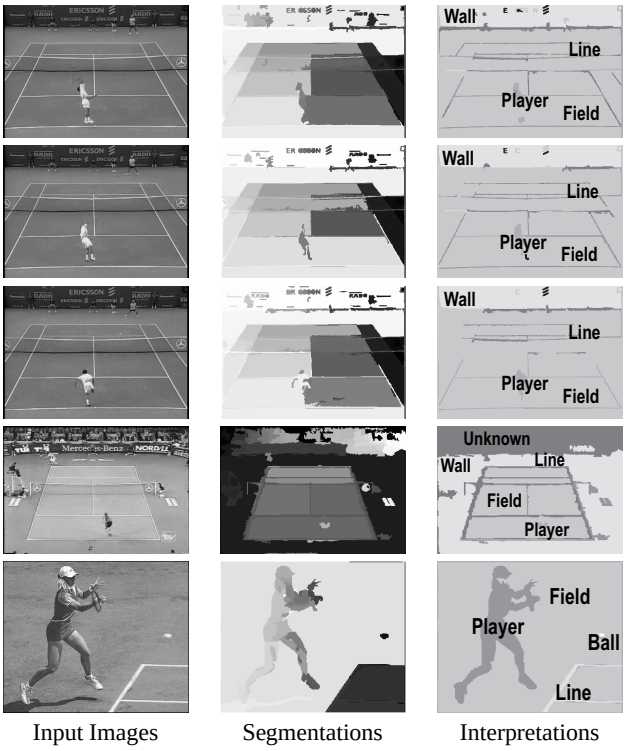

Fig. 2. Tennis domain results

8. 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] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, 1983.

[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] 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.

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

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