# AN ONTOLOGY INFRASTRUCTURE FOR MULTIMEDIA REASONING

*N. Simou*[1], *C. Saathoff*[3], *S. Dasiopoulou*[2],*E. Spyrou*[1], *N. Voisine*[2], *V. Tzouvaras*[1], *I. Kompatsiaris*[2], *Y. Avrithis*[1],*and S. Staab*[3]

[1]Image, Video and Multimedia Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, GR-15773 Zographou, Athens, Greece
[2]Informatics and Telematics Institute / Centre for Research and Technology Hellas, GR 57001 Thermi-Thessaloniki, Greece
[3]University of Koblenz, Institute for Computer Science, D-56016 Koblenz, Germany

## ABSTRACT

In this paper, an ontology infrastucture for multimedia reasoning is presented, making it possible to combine low-level visual descriptors with domain specific knowledge and subsequently analyze multimedia content with a generic algorithm that makes use of this knowledge. More specifically, the ontology infrastructure consists of a domain-specific ontology, a visual descriptor ontology (VDO) and an upper ontology. In order to interpret a scene, a set of atom regions is generated by an initial segmentation and their descriptors are extracted. Considering all descriptors in association with the related prototype instances and relations, a genetic algorithm labels the atom regions. Finally, a constraint reasoning engine enables the final region merging and labelling into meaningful objects.

## 1. INTRODUCTION

Recently, there is a growing research interest in the extraction of high-level semantic concepts from images and video using low-level multimedia features and domain knowledge. Significant progress has been made on automatic segmentation or structuring of multimedia content and the extraction of low-level features within such content [1]. However, comparatively little progress has been made on interpretation and generation of semantic descriptions of visual information. More importantly, most analysis techniques focus on specific application domains, making it hard to generalize in case other domains need to handled.

Due to the limitations of the state of the art multimedia analysis systems [2], it is acknowledged that in order to achieve semantic analysis of multimedia content, ontologies [3] are essential to express semantics in a formal machine-processable representation. Ontology-based metadata creation currently addresses mainly textual resources or simple annotation of photographs [4]. In well-structured applications (e.g. sports and news broadcasting) domain-specific features that facilitate the modelling of higher level semantics can be extracted [5]. A priori knowledge representation models are also used to assist semantic-based classification and clustering [6]. However, most such techniques are either not suitable for multimedia content analysis, or too correlated with the specific domains they are designed for.

In [7] a novel framework for video content understanding that uses rules constructed from knowledge bases and multimedia ontologies is presented. In [8], multimedia ontologies are semi-automatically constructed using a data-driven approach. [9] presents automatic techniques for extracting semantic concepts and discovering semantic relations among them and evaluates several techniques for visual feature descriptors extraction. In [10], semantic entities, in the context of the MPEG-7 standard, are used for knowledge-assisted video analysis and object detection, while in [11] MPEG-7 compliant low-level descriptors are mapped to intermediate-level descriptors forming an object ontology. It is evident from all such approaches, that a generic multimedia content analysis framework is required that makes use of knowledge stored in multimedia-enabled ontologies.

The framework presented in this paper combines low-level visual descriptors and domain-specific knowledge represented in an ontology infrastructure with a generic analysis scheme to semantically interpret and annotate multimedia content. The infrastructure consists of (i) a domain-specific ontology that provides the necessary conceptualizations for the specific domain, (ii) multimedia ontologies that model the multimedia layer data in terms of low level features and media structure descriptors, and (iii) a core on-

This research was partially supported by the European Commission under contract FP6-001765 aceMedia.

tology (DOLCE) that bridges the previous ontologies in a single architecture. During image/video analysis, a set of atom-regions is generated by an initial segmentation, and MPEG-7 visual descriptors are extracted for each region. A distance measure between these descriptors and the ones of the prototype instances included in the domain ontology is estimated using a neural network approach. A genetic algorithm then decides the initial labelling of the atom regions with a set of hypotheses, where each hypothesis represents a class from the domain ontology. Finally, a constraint reasoning engine enables the final merging of the regions, while at the same time reducing the number of hypotheses. This approach is generic and applicable to any domain as long as new domain ontologies are designed and made available.

The remainder of the paper is structured as follows: section 2 describes the ontology infrastructure. Section 3 describes the Genetic Algorithm approach, while section 4 describes the proposed reasoning engine. Results are presented in section 5 and conclusions are drawn in section 7.

## 2. ONTOLOGY INFRASTRUCTURE

There are two main factors that breed the need for a knowledge infrastructure for multimedia analysis. Firstly the fact that reasoners have to deal with large numbers of instantiations of the concepts an properties defined in ontologies, in cases of reasoning with multimedia data on large scale, and secondly that multimedia data comes in two separate though intertwined layers, multimedia and content layer. The multimedia layer deals with the semantics of properties related to the representation of content within the media data itself while on the other hand the content layer deals with the semantics of the actual content contained in the media data as it is perceived by the human media consumer.

Hence the knowledge infrastructure should model the multimedia layer data so as to support extraction and inferencing of content layer data. The ontology infrastructure used integrates these two layers consisting of:

- *Multimedia ontologies* that model the multimedia layer data in terms of low level features and media structure descriptors, namely the *Visual Descriptors Ontology* (VDO), based on an RDF representation of the MPEG-7 Visual Descriptors, and the *Multimedia Structure Ontology* (MSO), based on the MPEG-7 MDS.
- *Domain Ontologies* that provide the necessary conceptualizations of the content layer, for a specific application domain.
- A *Core Ontology* that models primitives at the root of the concept hierarchy and can be exploited by both types of ontologies. It is also meant to bridge between the other ontologies within the architecture.

The knowledge infrastructure is set up using RDFS. This approach is expected to be complemented by using an appropriate sub-language of OWL at a later stage. This decision reflects that a full usage of the increased expressiveness of OWL requires specialized and more advanced inference engines, especially when dealing with large numbers of instances.

The Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) was explicitly designed as core ontology. The RDFS version of DOLCE currently contains about 79 high level concepts and 81 high level properties among them. DOLCE contains explicit conceptualizations by including the concept of qualities that can be perceived, as well as spatio-temporal concept descriptions. However, reasoning with spatio-temporal descriptions requires the coding of additional relations that describe the relationship between space and/or time regions. Based on concepts taken from Region Connecting Calculus, Allen's interval calculus and directional models, we have carefully extended DOLCE to accomodate the corresponding directional and topological relationships in the spatial and temporal domains.

The top-level multimedia content entities of the MSO are described in MPEG-7 Multimedia Description Schemes (MDS) FCD. Within MPEG-7, multimedia content is classified into five types: Image, Video, Audio, Audiovisual and Multimedia. Each of these types has its own segment subclasses. The Segment DS describes a spatial and/or temporal fragment of multimedia content. A number of specialized subclasses are derived from the generic Segment DS. These subclasses describe the specific types of multimedia segments, such as video segments, moving regions, still regions and mosaics, which result from spatial, temporal and spatiotemporal segmentation of the different multimedia content types. Multimedia resources can then be accordingly decomposed into sub-segments through spatial, temporal, spatiotemporal or media source decomposition.

The VDO contains a set of visual descriptors to be used for knowledge-assisted analysis of multimedia content. By the term descriptor we mean a specific representation of a visual feature (color, shape, texture etc) that defines the syntax and the semantics of a specific aspect of the feature (dominant color, region shape etc). The entire VDO follows closely the specification of the MPEG-7 Visual Part, but several modifications were carried out in order to adapt to the datatype representations available in RDFS.

In order to extract a set of prototype low-level visual descriptors for different domain concepts and integrate them into the ontology structure, it must be clear how domain concepts can be linked with actual instance data without having to cope with meta-modelling. For this purpose, we have enriched the knowledge base with instances of domain

concepts that serve as *prototypes* for these concepts. Each of these is linked to the appropriate visual descriptor instances.

## 3. KNOWLEDGE-ASSISTED ANALYSIS

The domain ontology represents the required knowledge for interpreting each image or video scene, which is a mapping of image regions to the corresponding domain-specific semantic definition. Classes within the ontology have been defined to represent the different types of visual information while subclasses represent the different ways to calculate a visual feature. Each real-world object is allowed to have more than one instantiations. Currently, three spatial relations and three low-level descriptors are supported. These descriptors are: adjacency **(**$ADJ$**)**, below **(**$BEW$**)**, and inclusion **(**$INC$**)** relations, and dominant color **(**$DC$**)**, motion **(**$MOV$**)** and compactness **(**$CPS$**)** descriptors. Enriching the ontology with domain specific knowledge results in populating it with appropriate instances, i.e. prototypes for the objects to be detected.

During preprocessing, color segmentation [12][1]) and motion segmentation [13][11]) are combined to generate a set of over-segmented atom-regions. The extraction of the low-level descriptors for each atom-region is performed using the MPEG-7 eXperimentation Model(XM) [1]. Motion estimation is based on 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 each region. Global motion compensation is based on estimating the $8$ parameters of the bilinear motion model for camera motion, using an iterative rejection procedure [14]. Finally, the compactness descriptor is calculated by the area and the perimeter of the region.

After preprocessing, 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. A genetic algorithm is used to overcome the computational time constraints of testing all possible configurations [15]. In this approach, each individual represents a possible interpretation of the examined scene, i.e the identification of all atom regions. 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.

The degree of matching between regions, in terms of low-level visual and spatial features respectively, is defined as:

- 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 $om_j$, to provide an estimation of the degree of matching between an object model $om_j$ and a region $R_i$. $\mathcal{I}_M(R_i, om_j)$ 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.

The employed fitness function that considers the above matching estimations for all atom-regions is defined as:

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

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. Any neighboring regions belonging to the same object according to the generated optimal solution are simply merged. For each object that fails to comply the concept of unknown object is introduced.

Our approach to implement the interpretation function $\mathcal{I}_M$ used for the fitness function is based on a back-propagation neural network. When the task is to compare two regions based on a single descriptor, several distance functions can be used; however, there is not a single one to include all descriptors with different weight on each. This is a problem that the neural network handles. Its input consists of the low-level descriptions of both of an atom region and an object model, while its response is the estimated normalized distance between the atom region and the model. 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. This distance is then used for the interpretation function $\mathcal{I}_M$.

## 4. CONSTRAINT REASONING ENGINE

The analysis procedure described in section 3 results in an image segmented into a number of atom regions, each labeled with an initial set of hypotheses. Each hypothesis corresponds to one object description defined in the domain ontology. Although at this stage the atom-regions bear semantic information, further processing is required to derive a segmentation where each segment represents a meaningful object. To accomplish this, the limitations posed by the numerically based segmentation algorithms need to be overcome, i.e. atom-regions corresponding to only part instead

of the complete object, loss of object connectivity etc. In the following we describe an approach to meet this requirements based on reasoning on the labels and spatiotemporal information of the considered atom-regions.

The input of the proposed reasoning system consists of the set of atom-regions along with their initial labels as resulted following the former analysis procedure. The corresponding output is a reduced number of atom-regions, which coincide with real objects, within a plausible degree of accuracy, and a reduced set of hypotheses for each atom-region. The reasoning process is based on the extracted labels and spatiotemporal features of the examined atom-regions in association with the information included in the domain ontology.

The integration of low-level features in the reasoning process further improves the plausibility of the detection results. For example, a merging indicated by the defined rules should be performed only if the shape of the resulting segment conforms to the shape of one of the plausible object classes corresponding to the merged atom-regions. Obviously, this raises the need for incorporation of low-level feature matching into the reasoning system, which on the one hand can lead to computational problems and on the other hand reduces the number of eligible reasoning systems, because means to extend the system must be available.

**Fig. 1**. Input image.

This can be better understood through the example of Fig. 1, which is initially segmented and labelled as illustrated in Fig. 2. Regions labelled as 'sky', 'field' and 'mountain' are expected to be merged. Furthermore, more complex regions such as 'roof' and 'wall' are evaluated in association with each other. In other words, since a region has bright red color and geometrically fits to the description given for a roof it may be labelled as 'roof'. On the other hand, a white rectangle is difficult to be assigned a label alone, due to its very general features appearing in a number of prototype instances; however, according to available spatiotemporal information (white rectangle below roof) the 'wall' label is assigned to it.

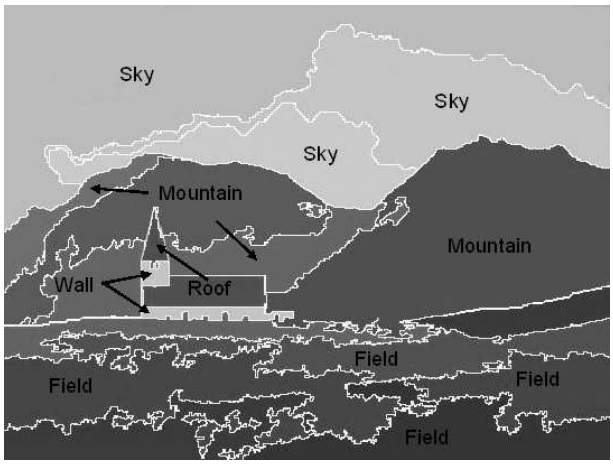


**Fig. 2**. Image after initial segmentation and labelling.

The whole process is iterative, as the actual region merging cannot be implemented efficiently within a reasoning system. Thus, the reasoner identifies regions that are to be merged, by adding a relation between them. Such relations are interpreted within a second step, where regions are merged, and any associated visual descriptors and relations are updated. New region labels are estimated, and hypotheses are then constructed. The output of this analysis step again serves as input for the reasoner in an iterative fashion until a stable state is reached, i.e. no new information can be inferred.

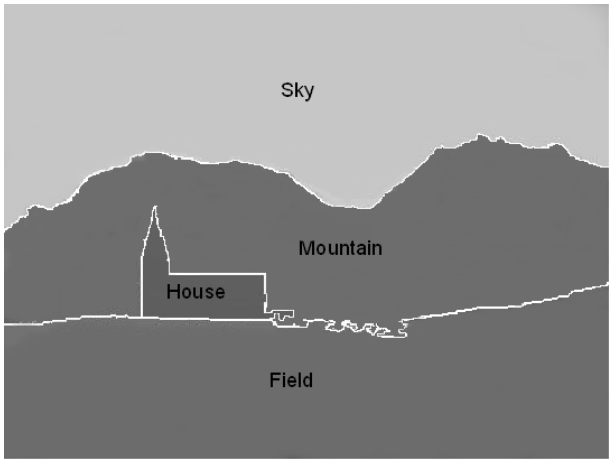


**Fig. 3**. Output of the constraint reasoner.

Using this approach, objects with similar visual characteristics can be discriminated in terms of their spatiotemporal behavior and the visual context on which they occur.

Furthermore, based on the output of the described reasoning process, further analysis becomes feasible, aiming at the generation of higher-level semantics, such as recognition of complex objects or events, which cannot be represented in terms of their visual features. In our example this is shown at Fig. 3. The 'sky', 'field' and 'mountain' regions have been merged but also regions 'wall' and 'roof' have been merged with the label 'house' assigned to the resulting region.

## 5. RESULTS

The presented ontology-based framework was used to extract semantic descriptions of a variety of MPEG-2 videos of the Formula One and Tennis domains. The corresponding domain ontologies, i.e the defined object classes along with their low-level features and spatial interrelations are illustrated in Table 1. A training set of manually annotated videos was used to populate the domain ontologies with prototype instances.

| Object Class | Low-level descriptors | 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 1**. Formula One and Tennis domain definitions.

As illustrated in Fig. 4 and 5, the system output is a segmentation mask outlining the semantic description of the scene where different colors representing the object classes defined in the domain ontology are assigned to the generated atom-regions.

Excluding the process of motion information extraction, the required analysis time was between 5 and 10 seconds per frame. The use of spatial information captures part of the visual context, consequently resulting in the extraction of more meaningful descriptions provided that the initial color-based segmentation did not segment two objects as one atom-region.

## 6. CONCLUSION

In this paper an approach is described that combines multimedia domain knowledge, a knowledge-assisted analysis and a constraint reasoner in order to extract high-level semantic knowledge from images and video. The developed ontology infrastructure efficiently relates domain knowledge with the semantics of the visual part of MPEG-7 through an upper harmonizing ontology. After a preprocessing step, a genetic algorithm generates a set of region label hypotheses, which are then processed by a constraint reasoning engine in an iterative fashion to enable the final region merging and labelling. The entire approach is generic, in the sense that all domain-specific information solely resides in the domain ontology; the same analysis framework has been tested on two different domains simply by switching the associated domain ontology, with promising initial results. This research is ongoing and future work includes implementation of larger scale domain ontologies enhanced with multimedia descriptions, relations and rules to evaluate the proposed methodology on a large set of multimedia content.

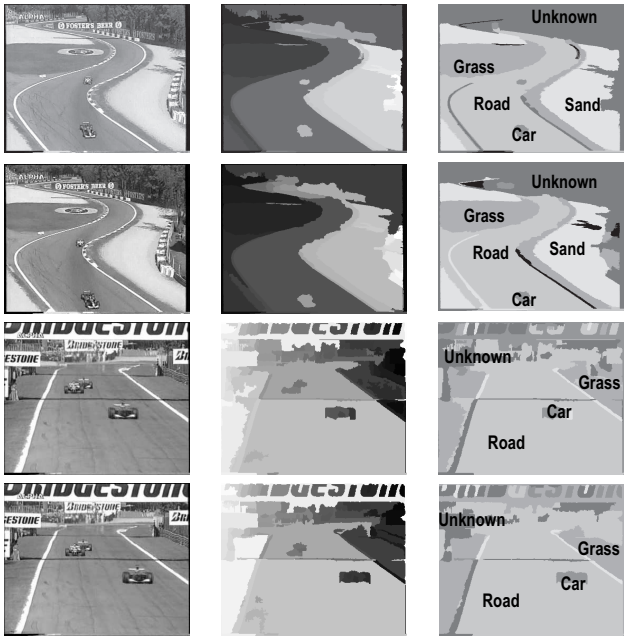


**Fig. 4**. Formula One domain results.

## 7. REFERENCES


[1] 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.

[2] O. Mich R. Brunelli and C.M. Modena. A survey on video indexing. *Journal of Visual Communications and Image Representation*, 10:78–112, 1999.

[3] Steffen Staab and Rudi Studer. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer Verlag, Heidelberg, 2004.

[4] J. Wielemaker A.Th. Schreiber, B. Dubbeldam and B.J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, May/June 2001.

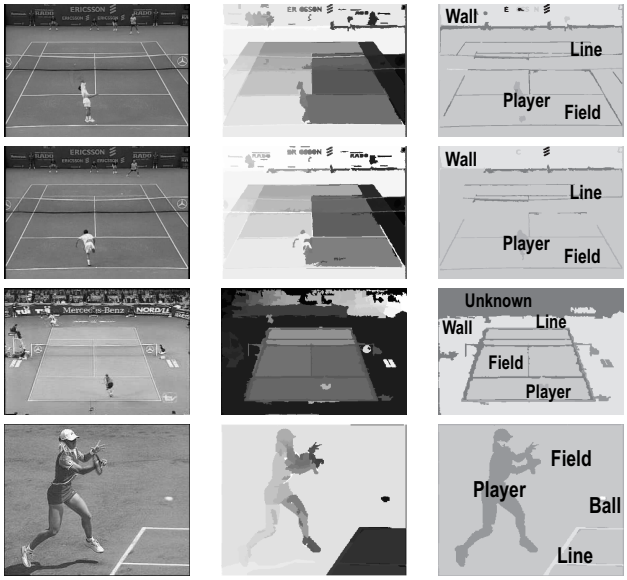


**Fig. 5**. Tennis domain results.


[5] 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.

[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] Belle Tseng Alejandro Jaimes and John R. Smith. In *Proc. IEEE International Conference on Image and Video Retrieval (ICME 2003)*.

[8] Alejandro Jaimes and John R. Smith. In *Proc. IEEE International Conference on Multimedia and Expo (ICME 2003)*.

[9] Ana B. Benitez and Shih-Fu Chang. In *Proc. IEEE International Conference on Image and Video Retrieval (ICME 2002)*.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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