A Learning Approach to Semantic Image Analysis

G. Th. Papadopoulos*†, P. Panagi*†, S. Dasiopoulou*†, V. Mezaris† and I. Kompatsiaris†
*Information Processing Laboratory
Electrical and Computer Engineering Department
Aristotle University of Thessaloniki, Greece

†Informatics and Telematics Institute
1st Km Thermi-Panorama Road
Thessaloniki, GR-57001 Greece
Email: {papad, panagi, dasiop, bmezaris, ikom}@iti.gr

***Abstract*— In this paper, a learning approach coupling Support Vector Machines (SVMs) and a Genetic Algorithm (GA) is presented for knowledge-assisted semantic image analysis in specific domains. Explicitly defined domain knowledge under the proposed approach includes objects of the domain of interest and their spatial relations. SVMs are employed using low-level features to extract implicit information for each object of interest via training in order to provide an initial annotation of the image regions based solely on visual features. To account for the inherent visual information ambiguity, fuzzy spatial relations along with the previously computed initial annotations are supplied to a genetic algorithm, which decides on the globally most plausible annotation. Experiments with images of the beach vacation domain demonstrate the performance of the proposed approach.**

I. Introduction

Recent advances in both hardware and software technologies have resulted in an enormous increase of the images that are available in multimedia databases or over the internet. As a consequence, the need for techniques and tools supporting their effective and efficient manipulation has emerged. To this end, several approaches have been proposed in the literature regarding the tasks of indexing, searching and retrieval of images [1][2].

The very first attempts to address these issues concentrated on visual similarity assessment via the definition of appropriate quantitative image descriptions, which could be automatically extracted, and suitable metrics in the resulting feature space. Coming one step closer to treating images the way humans do, these were later adapted to a finer granularity level, making use of the output of segmentation techniques applied to the image [1]. Whilst low-level descriptors, metrics and segmentation tools are fundamental building blocks of any image manipulation technique, they evidently fail to fully capture by themselves the semantics of the visual medium; achieving the latter is a prerequisite for reaching the desired level of efficiency in image manipulation. To this end, research efforts have concentrated on the semantic analysis of images, combining the aforementioned techniques with *a priori* domain specific knowledge, so as to result in a high-level representation of images [2]. Domain specific knowledge is utilized for guiding low-level feature extraction, higher-level descriptor derivation and symbolic inference.

Depending on the adopted knowledge acquisition and representation process, two types of approaches can be identified in the relevant literature: implicit, realized by machine learning methods, and explicit, realized by model-based approaches. The usage of machine learning techniques has proven to be a robust methodology for discovering complex relationships and interdependencies between numerical image data and the perceptually higher-level concepts. Moreover, these elegantly handle problems of high dimensionality. Among the most commonly adopted machine learning techniques are Neural Networks (NNs), Hidden Markov Models (HMMs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Genetic Algorithms (GAs) [4][5]. On the other hand, model-based image analysis approaches make use of prior knowledge in the form of explicitly defined facts, models and rules, i.e. they provide a coherent semantic domain model to support "visual" inference in the specified context [6][7].

In this paper, a semantic image analysis approach is proposed that combines two types of learning algorithms, namely SVMs and GAs, with explicitly defined knowledge in the form of an ontology that specifies domain objects and fuzzy spatial relations. SVMs are employed for performing an initial mapping between low-level visual features and the domain objects in the ontology (i.e. generating an initial hypothesis set for every image region) at the region level, whereas a GA is subsequently used to optimize this mapping over the entire image, while taking into account spatial relations. Application of the proposed approach to images of the specified domain results in the generation of fine granularity semantic representations, i.e. a segmentation map with semantic labels attached to each segment, by employing high level reasoning techniques. These initial labels can be used to infer additional knowledge.

The paper is organized as follows: Section II presents the overall system architecture. Sections III and IV describe the employed low- and high-level knowledge respectively. Sections V and VI detail individual system components. Experimental results for the beach vacation domain are presented in Section VII and conclusion are drawn in Section VIII.

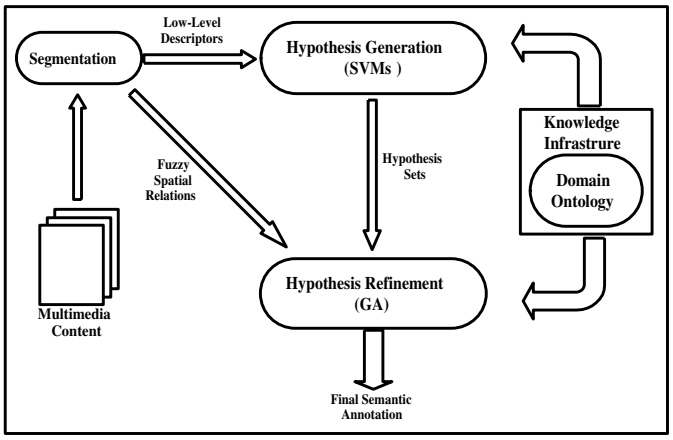

Fig. 1. System Architecture

II. System Overview

The overall architecture of the proposed system for semantic image analysis is illustrated in Fig. 1. Initially, a segmentation algorithm is applied in order to divide the given image into regions, which are likely to represent meaningful semantic objects. Then, for every resulting segment, low-level descriptions and spatial relations are estimated, the latter according to the relations supported by the domain ontology.

Estimated low-level descriptions for each region are employed for generating initial hypotheses regarding the region's semantic label. This is realized by evaluating the compound low-level descriptor vector by a set of SVMs, each trained to identify instances of a single concept defined in the ontology. SVMs were selected for this task due to their generalization ability and their efficiency in solving high-dimensionality pattern recognition problems [8][9].

The generated hypothesis sets for each region with the associated degrees of confidence for each hypothesis along with the spatial relations computed for every image segment, are subsequently employed for selecting a globally optimal set of semantic labels for the image regions by introducing them to a genetic algorithm. The choice of a GA for this task is based on its extensive use in a wide variety of global optimization problems [10], where they have been shown to outperform traditional methods, and is further endorsed by the authors previous experience [11], which showed promising results.

III. Low-level visual information processing

A. Segmentation and feature extraction

In order to implement the initial hypothesis generation procedure, the examined image has to be segmented into regions and suitable low-level descriptions have to be extracted for every resulting segment. In the current implementation, an extension of the Recursive Shortest Spanning Tree (RSST) algorithm has been used for segmenting the image [12].

Considering low-level descriptions, specific descriptors of the MPEG-7 standard have been selected, namely the *Scalable Color*, *Homogeneous Texture*, *Region Shape* and *Edge Histogram* descriptors. Their extraction is performed according to the guidelines provided by the MPEG-7 eXperimentation Model (XM) [13]. The above descriptors are extracted for every computed image segment and are combined in a single feature vector. This vector constitutes the input to the SVMs framework which computes the initial hypothesis set for every segment, as will be described in Section V.

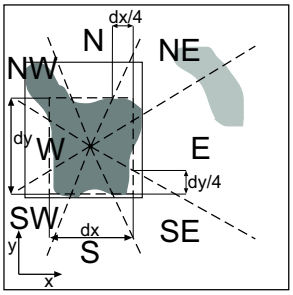

Fig. 2. Fuzzy directional relations definition

B. Fuzzy spatial relations extraction

Exploiting domain-specific spatial knowledge in image analysis tasks is a common practice among the object recognition community. It is generally observed that objects tend to be present in a scene within a particular spatial context and thus spatial information can substantially assist in discriminating between objects exhibiting similar visual characteristics. Among the most commonly adopted spatial relations, directional ones have received particular interest. They are used to denote the order of objects in space. In the present analysis framework, eight fuzzy directional relations are supported, namely *North* (*N*), *East* (*E*), *South* (*S*), *West* (*W*), *South-East* (*SE*), *South-West* (*SW*), *North-East* (*NE*) and *North-West* (*NW*).

Fuzzy directional relations extraction in the proposed analysis approach builds on the principles of projection- and angle-based methodologies [14][15] and consists of the following steps. First, a *reduced box* is computed from the *ground* object's (the object used as reference and is pointed in dark grey in Fig. 2) Minimum Bounding Rectangle (MBR) so as to include the object in a more representative way. The computation of this *reduced box* is performed in terms of the MBR compactness value *c*, which is defined as the value of the fraction of the objects's area to the area of the respective MBR: If the initially computed *c* is below a threshold *T*, the ground objects's MBR is reduced repeatedly until the desired threshold is satisfied. Then, eight cone-shaped regions are formed on top of this reduced box, as illustrated in Fig. 2, each corresponding to one of the defined directional relations. The percentage of the *figure* object (whose relative position is to be estimated and is pointed in light grey in Fig. 2) points that are included in each of the cone-shaped regions determines the degree to which the corresponding directional relation is satisfied. After extensive experimentations, the value of threshold *T* was set equal to *0.85*.

IV. KNOWLEDGE INFRASTRUCTURE

Among the possible domain knowledge representations, ontologies [16] present a number of advantages, the most important being that they provide a formal framework for supporting explicit, machine-processable semantics definition and they enable the derivation of new knowledge through automated inference. Thus, ontologies are suitable for expressing multimedia content semantics so that automatic semantic analysis and further processing of the extracted semantic descriptions is allowed. Following these considerations, a domain ontology was developed for representing the knowledge components that need to be explicitly defined under the proposed approach. This contains the semantic concepts that are of interest in the examined domain (e.g. in the beach vacation domain: Sea, Sand, Person, etc.), as well as their spatial relations. The values of the latter for the concepts of the given domain, as opposed to concepts themselves that are manually defined, are estimated according to the following ontology population procedure:

Let $S = \{s_i \ , \ i = 1, ..., \ N\}$ denote the set of regions produced for an image by segmentation, $O = \{o_j \ , \ j = 1, ...\}$ denote the set of objects defined in the employed domain ontology and

$$R = \{r_k, \ k = 1, ..., K\} = \quad (1)$$

$$= \{ \ N, \ NW, \ NE, \ S, \ SW, \ SE, \ W, \ E\}, \quad (2)$$

denote the set of supported spatial relations. Then, the degree to which s_i satisfies relation r_k with respect to s_j can be denoted as $I_{r_k}(s_i, s_j)$, where the values of function I_{r_k} are estimated according to the procedure of Section III-B and belong to $[0, 1]$. To populate the ontology, this function needs to be evaluated over a set of segmented images with ground truth annotations, that serves as a training set. More specifically, the mean values, $I_{r_k mean}$, of I_{r_k} are estimated, for every k over all region pairs of segments assigned to objects (o_i, o_j), $i \neq j$, and are stored in the ontology. These constitute the constraints input to the optimization problem which is solved by the genetic algorithm, as will be described in Section VI.

V. INITIAL HYPOTHESIS GENERATION

As already described in Section II, a Support Vector Machines (SVMs) structure is utilized to compute the initial hypothesis set for every image segment. Specifically, an individual SVM is introduced for every defined concept of the employed domain ontology, to detect the corresponding instances. Each SVM is trained under the '*one-against-all*' approach. For that purpose, the training set assembled in Section IV is employed and the combined region feature vector, as defined in Section III-A, constitutes the input to each SVM. For the purpose of initial hypothesis generation, every SVM returns a numerical value in the range $[0, 1]$ which denotes the degree of confidence to which the corresponding segment is assigned to the concept associated with the particular SVM. The metric adopted is defined as follows: For every input feature vector the distance *D* from the corresponding SVM's separating hyperplane is initially calculated. This distance is positive in case of correct classification and negative otherwise. Then, a sigmoid function [17] is employed to compute the degree of confidence, *DOC*, as follows:

$$DOC = \frac{1}{1 + e^{-mD}} \quad , \quad (3)$$

where the slope parameter *m* is experimentally set. The pairs of all domain concepts and their respective degree of confidence comprise each segment's hypothesis set. The above SVM structure was realized using the SVM software libraries of [18].

VI. HYPOTHESIS REFINEMENT

As outlined in Section II, after the initial set of hypotheses is generated (Section V), based solely on visual features, and the fuzzy spatial relations are computed for every pair of image segments (Section III-B), a genetic algorithm (GA) is introduced to decide on the optimal image interpretation. The GA is employed to solve a global optimization problem, while exploiting the available domain spatial knowledge, and thus overcoming the inherent visual information ambiguity. Spatial knowledge is obtained as described in Section IV and the resulting learnt fuzzy spatial relations serve as constraints denoting the "allowed" domain objects spatial topology.

Fuzzy spatial constraints verification factor.

Let

$$I_M(g_{ij}) \ \equiv \ DOC_{ij} \ , \quad (4)$$

denote the degree to which the visual descriptors extracted for segment s_i match the ones of object o_j, where g_{ij} represents the particular assignment of o_j to s_i. Thus, $I_M(g_{ij})$ gives the degree of confidence, DOC_{ij}, associated with each hypothesis and takes values in the interval $[0, 1]$.

Then, the function $I_S \ (g_{ij}, \ g_{pq})$ is defined as one that returns the degree to which the spatial constraint between the g_{ij}, g_{pq} object to segments mappings is satisfied. $I_S \ (g_{ij}, \ g_{pq})$ is set to receive values in the interval $[0, 1]$, where '1' denotes an allowable relation and '0' denotes an unacceptable one, based on the learnt spatial constraints. To calculate this value the following procedure is used:

Let $I^{'}_{r_k}(s_u, \ s_v)$ denote the degrees to which each spatial relation is verified for a certain pair of segments $s_u, \ s_v$ of the examined image and $o_s, \ o_t$ denote the domain defined concepts assigned to them respectively. A normalized euclidean distance $d(g_{us}, g_{vt})$ is calculated, with respect to the corresponding spatial constraint, as introduced in Section IV, based on the following equation:

$$d(g_{us}, g_{vt}) = \frac{\sqrt{\sum_{k=1}^{8} (I_{r_k mean}(o_s, o_t) - I^{'}_{r_k}(s_u, s_v))^2}}{\sqrt{8}} \quad , \quad (5)$$

which receives values in the interval $[0, 1]$. The function $I_S \ (g_{us}, \ g_{vt})$ is then defined as:

$$I_S\ (g_{us},\ g_{vt})\ =\ 1 - d(g_{us}, g_{vt}) \tag{6}$$

and takes values in the interval $[0, 1]$ as well.

Implementation of genetic algorithm.

As has already been described, the proposed algorithm uses as input the initial hypothesis sets (generated by the SVMs structure), the fuzzy spatial relations extracted between the examined image segments, and the spatial-related domain knowledge as produced by the particular training process. Under the proposed approach, each chromosome represents a possible solution. Consequently, the number of the genes comprising each chromosome equals the number N of the segments s_i produced by the segmentation algorithm and each gene assigns a defined domain concept to an image segment.

A population of 200 chromosomes is employed, and it is initialized with respect to the input set of hypotheses. An appropriate *fitness function* is introduced to provide a quantitative measure of each solution fitness, i.e. to determine the degree to which each interpretation is plausible:

$$f(C)\ =\ \lambda \times FS_{norm}\ +\ (1-\lambda) \times SC_{norm}\ , \tag{7}$$

where C denotes a particular chromosome, FS_{norm} refers to the degree of low-level descriptors matching, and SC_{norm} stands for the degree of consistency with respect to the provided spatial domain knowledge. The variable λ is introduced to adjust the degree to which visual features matching and spatial relations consistency should affect the final outcome. After thorough experimentation, λ was set to 0.35, which points out the importance of spatial context.

The values of SC_{norm} and FS_{norm} are computed as follows:

$$FS_{norm}\ =\ \frac{\sum_{i=1}^{N} I_M(g_{ij}) - I_{min}}{I_{max} - I_{min}}\ , \tag{8}$$

where $I_{min} = \sum_{i=1}^{N} \min_j I_m(g_{ij})$ is the sum of the minimum degrees of confidence assigned to each region hypotheses set and $I_{max}\ =\ \sum_{i=1}^{N} max_j I_m(g_{ij})$ is the sum of the maximum degrees of confidence values respectively.

$$SC_{norm}\ =\ \frac{\sum_{l=1}^{W} I_{S_l}(g_{ij}, g_{pq})}{W}\ , \tag{9}$$

where W denotes the number of the constraints that had to be examined.

After the population initialization, new generations are iteratively produced until the optimal solution is reached. Each generation results from the current one through the application of the following operators.

- Selection: a pair of chromosomes from the current generation are selected to serve as parents for the next generation. In the proposed framework, the Tournament Selection Operator [19], with replacement, is used.
- Crossover: two selected chromosomes serve as parents for the computation of two new offsprings. Uniform crossover with probability of 0.7 is used.
- Mutation: every gene of the processed offspring chromosome is likely to be mutated with probability of 0.008. If mutation occurs for a particular gene, then its corresponding value is modified, while keeping unchanged the degree of confidence.

To ensure that chromosomes with high fitness will contribute to the next generation, the overlapping populations approach was adopted. More specifically, assuming a population of m chromosomes, m_s chromosomes are selected according to the employed selection method, and by application of the crossover and mutation operators, m_s new chromosomes are produced. Upon the resulting $m + m_s$ chromosomes, the selection operator is applied once again in order to select the m chromosomes that will comprise the new generation. After experimentation, it was shown that choosing $m_s = 0.4m$ resulted in higher performance and faster convergence. The above iterative procedure continues until the diversity of the current generation is equal to/less than 0.001 or the number of generations exceeds 50.

VII. Experimental results

In this section, we present experimental results from testing the proposed approach in the domain of beach vacation images. First, a domain ontology had to be developed to represent the domain objects of interest and their relations. Four concepts are currently supported, namely *Sky*, *Sea*, *Sand* and *Person*.

Then, a set of 40 randomly selected images belonging to the beach vacation domain were used to assemble a training set for the low-level implicit knowledge acquisition (SVMs training) and computation of the fuzzy spatial constraints. A corresponding set of 400 images was similarly formed to serve as a test set for the evaluation of the proposed system performance. Each image of the training/test set was manually annotated according to the domain ontology definitions.

According to the SVMs training process, a set of instances were selected for every defined domain concept from the assembled training image set. The Gaussian radial basis function was used as a kernel function by each SVM, to allow for nonlinear discrimination of the samples. The low-level combined feature vector, as described in detail in Section III-A, is composed of 433 values, normalized in the interval $[-1, 1]$. On the other hand, for the acquisition of the fuzzy spatial constraints, the procedure described in Section IV was followed for each possible combination of the defined domain objects that were present in the employed image training set.

Based on the trained SVMs structure, initial hypotheses are generated for each image segment as described in Section V, which are then passed in the genetic algorithm along with the fuzzy spatial constraints in order to determine the globally optimal interpretation. In Fig. 3 indicative results are given showing the input image, the annotation resulting from the initial hypotheses set, considering for each image segment the

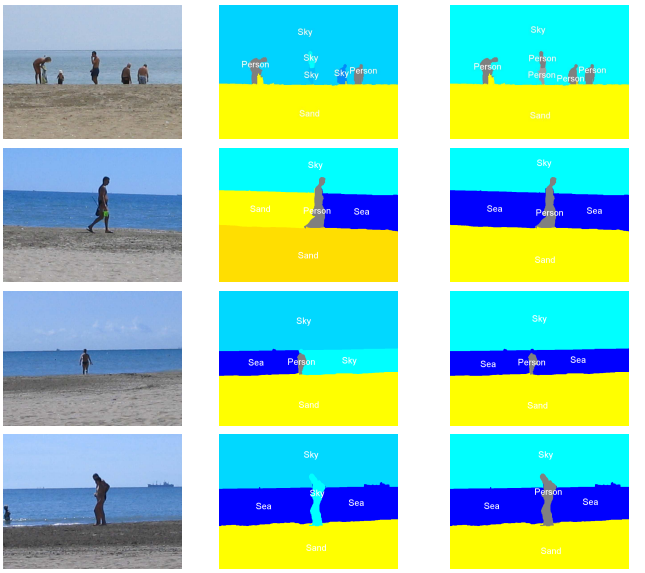

Fig. 3. Exemplar results for the beach domain.

TABLE I

NUMERICAL EVALUATION FOR THE BEACH VACATION DOMAIN.

	Proposed Approach				Method proposed in [3]	
	Initial Hypothesis		Final Interpretation		Final Interpretation	
Object	precision	recall	precision	recall	precision	recall
Sky	58.73%	92.50%	81.25%	97.50%	56.34%	98.92%
Sea	89.47%	53.13%	92.06%	60.42%	92.75%	66.67%
Sand	76.79%	97.73%	87.76%	97.73%	85.42%	93.18%
Person	72.34%	82.92%	70.91%	95.12%	78.38%	70.73%
Accuracy	75.95%		83.20%		77.48%	

hypothesis with the highest degree of confidence, and the final interpretation after the application of the genetic algorithm.

In Table I, quantitative performance measures are given in terms of precision and recall. We compared the performance of our method with the method described in [3]. It must be noted that for the numerical evaluation, any object present in the examined image test set that was not included in the domain ontology concept definitions, e.g. umbrella, was not taken into account.

After a careful observation of the presented results, we can confirm the good generalization ability of SVMs, regardless of the usage of a limited training set. Furthermore, we can justify the choice of using a genetic algorithm to reach an optimal image interpretation given degrees of confidence for visual similarity and spatial consistency against the domain definitions.

VIII. CONCLUSIONS

In this paper, an approach to knowledge-assisted semantic image analysis that couples Support Vector Machines (SVMs) with a Genetic Algorithm (GA) is presented. The proposed system was tested for the beach vacation domain and produced satisfactory results. Furthermore, the system can be easily applied to additional domains, given the fact that an appropriate domain ontology is defined and the corresponding training sets are formed.

ACKNOWLEDGMENT

The work presented in this paper was partially supported by the European Commission under contracts FP6-001765 aceMedia, FP6-027026 K-Space and COST 292 action.

REFERENCES

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, *"Content-based image retrieval at the end of the early years"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, iss. 12, pp. 1349-1380,2000.

[2] W. Al-Khatib, Y. F. Day, A. Ghafoor, P. B. Berra, *"Semantic Annotation of Images and Videos for Multimedia Analysis"*, 2nd European Semantic Web Conference (ESWC), Herakleion, Greece, 2005.

[3] K. Petridis, S. Bloehdorn, C. Saathoff, N. Simou, S. Dasiopoulou, V. Tzouvaras, S. Handschuh, Y. Avrithis, I. Kompatsiaris and S. Staab, *"Knowledge Representation and Semantic Annotation of Multimedia Content"*, IEE Proceedings on Vision Image and Signal Processing, Special issue on Knowledge-Based Digital Media Processing, Vol. 153, No. 3, pp. 255-262, June 2006.

[4] J. Assfalg, M. Berlini, A. Del Bimbo, W. Nunziat, P. Pala, *"Soccer Highlights Detection and Recognition using HMMs"*, IEEE International Conference on Multimedia & Expo (ICME), pp. 825-828, 2005.

[5] L. Zhang, F.Z. Lin, B. Zhang, *"Support Vector Machine Learning for Image Retrieval"*, International Conference on Image Processing, October, 2001.

[6] S. Dasiopoulou, V. Mezaris, V.K. Papastathis, I. Kompatsiaris, M.G. Strintzis, *"Knowledge-Assisted Semantic Video Object Detection"*, IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Analysis and Understanding for Video Adaptation, vol. 15, no. 10, pp. 1210-1224, 2005.

[7] L. Hollink, S. Little, J. Hunter, *"Evaluating the Application of Semantic Inferencing Rules to Image Annotation"*, 3rd International Conference on Knowledge Capture (K-CAP05), Banff, Canada, 2005.

[8] K. I. Kim, K. Jung, S. H. Park, and H. J. Kim, *"Support vector machines for texture classification"*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 11, pp. 1542.1550, Nov. 2002.

[9] O. Chapelle, P. Haffner and V. Vapnik, *"Support vector machines for histogram-based image classification"*, IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 10551064, September 1999.

[10] M. Mitchell, *"An introduction to genetic algorithms"*, MIT Press, 1995.

[11] N. Voisine, S. Dasiopoulou, F. Precioso, V. Mezaris, I. Kompatsiaris and M.G. Strintzis, *"A Genetic Algorithm-based Approach to Knowledge-assisted Video Analysis"*, Proc. IEEE International Conference on Image Processing (ICIP 2005), Genova, 2005.

[12] T. Adamek, N. O'Connor, N. Murphy, *"Region-based Segmentation of Images Using Syntactic Visual Features"*, Workshop on Image Analysis for Multimedia Interactive Services, (WIAMIS), Montreux, Switzerland, 2005.

[13] *"MPEG-7 Visual Experimentation Model (XM)"*, Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4062, Mar., 2001.

[14] S. Skiadopoulos, C. Giannoukos, N. Sarkas, P. Vassiliadis, T. Sellis, M. Koubarakis, *"2D topological and direction relations in the world of minimum bounding circles"*, IEEE Transactions on Knowledge and Data Engineering, vol. 17, iss. 12, pp. 1610-1623, 2005.

[15] Y. Wang, F. Makedon, J. Ford, L. Shen, D. Golding, *"Generating Fuzzy Semantic Metadata Describing Spatial Relations from Images using the R-Histogram"*, JCDL '04, June 7-11, Tucson, Arizona, USA, 2004.

[16] S. Staab and R. Studer, *"Handbook on ontologies"*, in Int. Handbooks on Information Systems. Berlin, Germany: Springer-Verlag, 2004.

[17] D. Tax and R. Duin, *"Using two-class classifiers for multi-class classification"*, in Proc. Int. Conf. Pattern Recognition, Quebec City, Canada, vol. 2, pp. 124127, 2002.

[18] C.-C. Chang and C.-J. Lin., *"LIBSVM: A library for support vector machines"*, http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[19] D. Goldberg, K. Deb, *"A comparative analysis of selection schemes used in genetic algorithms"*, In Foundations of Genetic Algorithms, G. Rawlins, 69-93, 1991.