A Comparative Study of Object-level Spatial Context Techniques for Semantic Image Analysis

G. Th. Papadopoulos, C. Saathoff, H. J. Escalante, V. Mezaris, I. Kompatsiaris and M. G. Strintzis

***Abstract*—In this paper, three approaches to utilizing object-level spatial contextual information for semantic image analysis are presented and comparatively evaluated. Contextual information is in the form of fuzzy directional relations between image regions. All techniques, namely a Genetic Algorithm (GA), a Binary Integer Programming (BIP) and an Energy-Based Model (EBM), are applied in order to estimate an optimal semantic image interpretation, after an initial set of region classification results is computed using solely visual features. Aim of this paper is the in-depth investigation of the advantages of each technique and the gain of a better insight on the use of spatial context. For this purpose, an appropriate evaluation framework, which includes several different combinations of low-level features and classification algorithms, has been developed. Extensive experiments on six datasets of varying problem complexity have been conducted for investigating the influence of typical factors (such as the utilized visual features, the employed classifier, the number of supported concepts, etc.) on the performance of each spatial context technique, while a detailed analysis of the obtained results is also given.**

I. Introduction

The extensive proliferation of multimedia capturing devices with high storage capabilities in combination with the continuously growing network access availability have resulted in the generation of literally vast image collections. The latter are being exchanged among individuals or are made available over the Internet. At the same time, common image manipulation tasks, like indexing, search and retrieval in such collections, often constitute an integral part of an individual's everyday activities at both personal or professional level. As a consequence, new needs have emerged regarding the development of advanced and user-friendly systems for the efficient manipulation of the image content [1]. For tackling these challenges, approaches that shift image processing to a semantic level have been proposed and so far exhibited promising results [2].

Among the approaches belonging to the latter category, semantic image analysis techniques, i.e. techniques aiming at localizing and recognizing the actual objects that are depicted in the image, have received particular attention. Their achievements and outcomes have been shown to significantly reinforce other image manipulation tasks, since they can provide a good foundation for boosting image classification [3], enabling the realization of complex queries [4] or facilitating further inference [5], to name a few. However, the efficiency of semantic image analysis approaches based on image segmentation and object recognition is significantly hindered by the ambiguity that is inherent in the visual medium. This is due to the fact that the localization and recognition of the real-world objects in unconstrained environments constitutes a challenging problem. For overcoming this limitation, among other solutions, the use of contextual information has been proposed [6].

Image context includes all possible information sources that can contribute to the understanding of the image content, complementarily to the use of the visual features. In the setting of semantic analysis, contextual information comprises any kind of relations between the semantic entities that can be present in an image (e.g. spatial, co-occurrence, scene-type information, etc.). Following the context acquisition procedure, contextual information can be used for: a) refining the image analysis results that have been computed based solely on visual features, by serving as a set of constraints that the former need to satisfy, and b) providing the appropriate prior knowledge that is required for performing inference and generating more detailed semantic descriptions. Out of the available contextual information types, spatial context is of increased importance in semantic image analysis. Spatial context represents and models the spatial configuration of the real-world objects and facilitates in discriminating between objects that exhibit similar visual characteristics.

Spatial contextual information can be divided into global- and local-level [7]. Global spatial context includes information about the overall spatial layout of the image and facilitates in identifying different scene configuration types. In [8], a framework is developed for modeling the correlations between the statistics of low-level features across the entire scene and the objects that it contains. Verbeek et al. [9] introduce a Conditional Random Field (CRF)-based scene labeling model that incorporates both local features and features aggregated over the whole image or large sections of it for performing semantic region labeling. In [10], a context-based vision system is proposed for place and object recognition, which relies on the principle of initially categorizing the image and subsequently using this information for providing contextual priors for the object recognition procedure. On the other hand,

G. Th. Papadopoulos and M. G. Strintzis are with the Informatics and Telematics Institute / Centre for Research and Technology Hellas, 6th Km Charilaou-Thermi Road, P.O.BOX 60361, Thermi 57001, Greece, and with the Electrical and Computer Engineering Department of Aristotle University of Thessaloniki, Greece.

C. Saathoff is with the WeST - Institute for Web Science and Technologies, University of Koblenz-Landau, Germany.

H. J. Escalante is with the Graduate Program in Systems Engineering of Universidad Autónoma de Nuevo León, Mexico.

V. Mezaris and I. Kompatsiaris are with the Informatics and Telematics Institute / Centre for Research and Technology Hellas, 6th Km Charilaou-Thermi Road, P.O.BOX 60361, Thermi 57001, Greece.

This work was supported by the European Commission under contracts FP7-248984 GLOCAL, FP7-214306 JUMAS and FP7-215453 WeKnowIt.

local spatial context concerns relations derived from the area that surrounds the object to be detected. The latter may include interactions between objects [11], [12], patches [13], [14] or pixels [15], [16], [17], [18]. In this work, spatial contextual information at object-level is considered.

Although a series of different and well-performing approaches to spatial context exploitation have been proposed [19], [20], [21], [22], the evaluation of each has been mostly limited to very few datasets (usually one or two) or rather specific application cases. On the other hand, a comprehensive study examining under which circumstances the use of spatial context is advantageous and how its resulting performance is affected by typical factors such as the utilized visual features, the employed classifier, the number of supported objects, different datasets of varying complexity, the amount of data used for spatial context acquisition or the number of regions that are present in the image, has not been performed.

In this paper, a comparative evaluation of three spatial context techniques for semantic image analysis is conducted with several different combinations of low-level features and classifiers on six datasets of varying problem complexity. Aim of this study is the in-depth investigation of the advantages of each spatial context approach and the gain of a better insight on the use of spatial contextual information. To achieve this, the three considered spatial context techniques, i.e. a Genetic Algorithm (GA), a Binary Integer Programming (BIP) and an Energy-Based Model (EBM), are selected so as to cover the main categories of the approaches that have been proposed in the literature. An appropriate evaluation framework, whose general structure is illustrated in Fig. 1, has been developed for realizing this study. Additionally, a novel quantitative measure, called Spatial Context Factor (SCF), is introduced for indicating the degree to which the spatial configuration of a given object is well-defined. As can be seen in Fig. 1, the examined image is initially segmented and two individual sets of visual features, namely MPEG-7 descriptors and SIFT-based features, are extracted for every resulting segment. In parallel, for every pair of image regions a corresponding set of fuzzy directional spatial relations are estimated. Then, each set of low-level features is in turn provided as input to three different classification algorithms, namely a Support Vector Machine (SVM), a Random Forest (RF) and a LogitBoost (LB). Each classifier aims at associating every region with a predefined high-level semantic concept based solely on visual information. The latter is used for denoting a real-world object that can be present in the examined image. Subsequently, the three aforementioned spatial context techniques, which perform on top of the initial classification results and follow different approaches for spatial context acquisition, are applied in order to estimate an optimal region-concept assignment. Extensive experiments have been conducted for investigating the influence of a series of factors on the performance of each spatial context technique, and a detailed analysis of the obtained results is given.

The paper is organized as follows: Section II presents an overview of the relevant literature and discusses the selection of the considered techniques. Section III outlines the visual information processing. The spatial relations extraction and the context acquisition procedure are detailed in Section IV. The selected spatial context exploitation techniques are described in Section V. Experimental results from the performed comparative evaluation as well as detailed analysis of them are presented in Section VI and conclusions are drawn in Section VII. The main symbols used in the remainder of the manuscript are outlined in Table I.

II. Object-level Spatial Context Techniques

Object-level spatial context approaches take into account information about the spatial configuration of the objects, in order to facilitate in their discrimination. These techniques can be roughly categorized using two main criteria: i) the complexity of the utilized contextual information and ii) the methodology followed for enforcing the acquired spatial constraints. With respect to the complexity of the spatial information, the following categories of methods have been proposed:

1a) methods examining adjacency characteristics: In [23], [14], [24], [9], a series of methods that take into account information about the adjacency between image regions for assigning the appropriate semantic concepts are proposed. Additionally, González-Díaz et al. [25] present a generative model that considers the length of the common boundary between pairs of regions. Examining the adjacency between image regions results in reduced expressiveness of the acquired contextual information, which in turn limits the use of this category of methods in specific application cases.

1b) approaches that make use of binary spatial relations: The methods of [26], [20], [27], [11], [28] follow a frequency counting approach for estimating spatial constraints between object pairs. Desai et al. [21] formulate the learning of a set of weights, which encode valid spatial configurations of individual object classes, as a convex optimization problem. Additionally, a maximum-likelihood approximation is followed for the acquisition of spatial contextual information in [13]. Saathoff et al. [29] use support and confidence as selection criteria for obtaining a set of binary constraints.

1c) methods supporting the use of fuzzy relations: A statistical learning approach to spatial context exploitation is described in [12], where fuzzy directional relations are considered and the impact of every acquired spatial constraint is adaptively adjusted. In [30], a fuzzy spatial relation ontology is developed for guiding image interpretation and facilitating the recognition of the semantic concepts it contains.

Regarding the methodologies followed for enforcing the acquired spatial constraints, these have been dominated by the use of Machine Learning (ML) and probabilistic techniques. The main categories that have been presented include:

2a) graphical modeling-based methods: A CRF-based approach is presented in [20] that incorporates both co-occurrence and spatial contextual information.

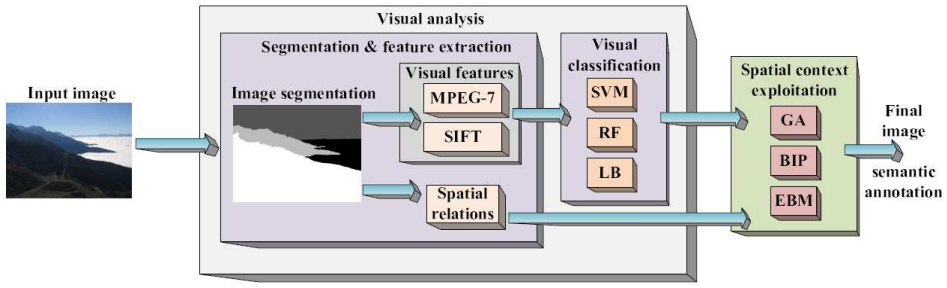

Fig. 1. Developed evaluation framework.

TABLE I
LEGEND OF MAIN SYMBOLS

Symbol	Description
$s_n,\ n \in [1, N]$	created image regions after segmentation
v_n	visual feature vector extracted for region s_n
$c_k,\ k \in [1, K]$	defined semantic concepts
$h_{nk} \equiv P(c_k \mid v_n)$	probability with which concept c_k is assigned to region s_n, using only visual features
$R = \{r_\gamma,\ \gamma \in [1, \Gamma]\}$	set of supported directional relations
$r_\gamma(s_n, s_m)$	degree of satisfaction of relation r_γ by the ordered region pair (s_n, s_m); this belongs to the continuous range $[0, 1]$
g_{nk}	assignment of concept c_k to region s_n, after spatial context exploitation
$freq(c_k)$	frequency of occurrence of concept c_k
$freq(c_k, c_l)$	co-occurrence frequency of concept pair (c_k, c_l)

CRFs are also used in [13] for encoding the objects' relative configuration and in [9] for incorporating spatial adjacency information during the assignment of high-level objects to local image patches. Carbonetto et al. propose a Markov Random Field (MRF)-model that combines image feature vectors with spatial relations for the task of object recognition in [14], while Heesch et al. [28] introduce a MRF with asymmetric Markov parameters to model the spatial and topological relationships between objects in structured scenes. Additionally, Bayesian Networks (BNs) are employed in the works of [11] and [19], for learning probabilistic spatial context models and for combining spatial context with visual and co-occurrence information, respectively. Torralba et al. [31] introduce the so called Boosted Random Fields (BRFs) for exploiting both local image data and spatial contextual information. Moreover, Yuan et al. [26] employ simple grid-structure graphical models to characterize the spatial dependencies between the objects depicted in the image. A tree graphical model is proposed to learn the spatial configuration of the different object categories in [22].

2b) statistical learning approaches: An extension of the original Latent Dirichlet Allocation (LDA) technique, in order to incorporate spatial information, is proposed in [32] for simultaneously segmenting and classifying objects that are present in the examined image. Similarly, extensions of the traditional probabilistic Latent Semantic Analysis (pLSA) technique are proposed in [33] and [25] for detecting different object categories and their approximate spatial layout, and for fusing local visual information with the global geometric layout of a segmented image, respectively. Additionally, a method termed Mutual Boosting is presented in [34] for incorporating spatial contextual information during object detection.

2c) methods that are based on optimization techniques and methods for solving systems of linear equations: Papadopoulos et al. [12] make use of a genetic algorithm for realizing image analysis as a global optimization problem, taking into account spatial contextual information. In [29], the problem of spatial context exploitation is formalized following a linear programming technique. Additionally, a spectral theory-based method is proposed in [24] for incorporating spatial information in the image labeling process. In [35], semantic image analysis is realized as an arc consistency checking problem with bilevel constraints, using qualitative spatial relations. Moreover, the exploitation of spatial contextual information between objects is formalized as a fuzzy constraint satisfaction problem in [36].

It must be noted that more elaborate approaches, which combine characteristics from more than one of the aforementioned categories, have also been proposed. For example, the GA method of [19] that is among the ones examined in this work follows a statistical learning approach for spatial constraints acquisition, while it makes use of a set of BNs for combining the spatial with the visual and the objects' co-occurrence information.

A. Selection of Considered Techniques

The core objective of this work is to gain a better insight and derive general observations regarding the use of object-level

spatial contextual information. For achieving this, three different techniques are considered. These are a Genetic Algorithm (GA), a Binary Integer Programming (BIP) and an Energy-Based Model (EBM). Specifically, the GA realizes image analysis as a global optimization problem [19]. This method incorporates a set of BNs for probabilistically adjusting the impact of the spatial, visual and concepts' co-occurrence information, while a statistical learning approach is followed for estimating complex fuzzy spatial constraints. The BIP also formalizes spatial context exploitation as an optimization problem and, in particular, it follows a linear programming methodology [29]. The latter technique makes use of binary spatial constraints, which are computed using support and confidence as selection criteria. The EBM represents the image as a fully connected graphical model [37], which in its current implementation is appropriately extended to include spatial information. Then, it estimates the objects' expected relative position, by calculating a set of fuzzy spatial constraints. All the aforementioned approaches make use of fuzzy directional relations.

The spatial context techniques that are included in the developed evaluation framework are selected so as to cover the main categories of the approaches that have been proposed in the literature. In particular, the EBM is representative of category ($2a$) described above, i.e. methods that associate every image region with a node in a graphical model that represents the image and, subsequently, the semantic image interpretation is estimated by performing inference in this model. Additionally, the EBM constitutes also an instance of category ($1c$), since it allows the use of fuzzy spatial constraints between the supported objects. On the other hand, the estimation and usage of a set of binary constraints classify the BIP to category ($1b$), while the formulation of region labeling as a linear programming problem is a typical case of class ($2c$). Moreover, the GA is in principle a member of ($2c$) and ($1c$), since it adopts a global optimization methodology for incorporating spatial information in the analysis process and supports the use of fuzzy spatial relations, respectively. However, the GA is also a member of ($2b$), since it follows a statistical learning approach for acquiring complex spatial contextual information and also employs a set of BNs for probabilistically adjusting the impact of the spatial versus the visual and the co-occurrence information.

In Table II, some of the most representative spatial context techniques of the literature are presented, where the type of the contextual information and the constraints enforcement procedure are given for every case. As can be seen, the methods of [25] and [32] utilize adjacency characteristics (i.e. the length of the regions' common boundary and the region adjacency property, respectively). This results in reduced expressiveness of the objects' spatial configuration. On the contrary, all selected spatial context techniques use fuzzy directional relations. Additionally, both the pLSA and LDA techniques, which are used by the methods of [25] and [32] respectively, frequently present overfitting problems and use approximations of the solution, especially when the network complexity is increased (i.e. large number of regions or objects). A series of techniques (i.e. the methods of [26], [20], [28] and [11]) follow a simple frequency counting approach of binary relations for spatial context acquisition. Although the aforementioned methods provide a more detailed representation of the objects' topology than the approaches of [25] and [32], the acquired spatial contextual information remains significantly simpler than the fuzzy constraints that the EBM and GA use. In parallel, a set of techniques rely on the use of graphical models (i.e. the methods of [26], [20], [28], [13] and [22]), where every image region is associated with a node of a model and spatial context exploitation is realized by performing inference in this model. The main drawback of these graphical modeling-based methods is that they often lead to intractable partition functions, especially when images with many regions are involved. The EBM is a representative of this category of techniques. However, EBMs are advantageous compared to other undirected graphical models that are widely used, like MRFs. This is mainly due to the fact that they allow the relaxation of the strict probabilistic assumptions and the avoidance of intractable partition functions [38]. Singhal et al. [11] make use of a series of BNs, which are gradually constructed and solved in an iterative manner, for approximating the final image interpretation. Nevertheless, the proposed greedy inference propagation scheme was shown to be significantly outperformed by the GA [12], which realizes image analysis as a global optimization problem. Moreover, the techniques of [13] and [22], apart from the limitations that derive from the usage of graphical models discussed above, utilize relatively simple spatial contextual information. More specifically, the method of [13] learns a set of weights for the employed binary spatial relations and the approach of [22] uses the height and the vertical position of the objects, while they both incorporate significant probabilistic assumptions for enabling the learning process. On the other hand, the technique of [21], beside the use of relatively simple spatial contextual information, adopts a greedy search algorithm for finding the optimal image interpretation. The latter characteristic constitutes a limitation when the problem complexity increases (i.e. a large number of objects or image regions is present), compared to the GA and the BIP that follow a global optimization approach. Furthermore, the BIP was experimentally shown to outperform the method of [36], although using the same methodology for acquiring the spatial constraints [29]. This is mainly due to the need of a monotonically decreasing objective function by the adopted fuzzy constraint satisfaction approach, similarly to other common optimization methods. The BIP, on the contrary, can use an arbitrary linear objective function, since it constitutes a particular type of linear programs. At this point, it must be highlighted that the selected GA technique presents a significant advantage over all the other methods presented already. This is that it incorporates a probabilistic approach for efficiently adjusting the impact of the available spatial, visual and co-occurrence information on the final outcome for every possible pair of objects. Taking into account all the aforementioned considerations, it is shown that the three selected spatial context techniques present advantageous characteristics, compared to other similar methods of the literature; hence, they are suitable for deriving general remarks on the use of spatial contextual information.

TABLE II
REPRESENTATIVE SPATIAL CONTEXT TECHNIQUES OF THE LITERATURE

Method	Context type	Constraints enforcement
[25]	Length of the common boundary between the regions	Extension of the traditional pLSA technique
[32]	Region adjacency	Latent topic model, extension of the traditional LDA technique
[26]	Frequency counting of binary relations	Grid-structure graphical models (2D-HMM, MRF, CRF)
[20]	Frequency matrices of pairwise relations	CRF
[28]	Frequency counting of pairwise relations between neighboring regions	MRF with asymmetric parameters
[11]	Frequency counting of binary relations	Set of BNs solved iteratively
[13]	ML approximation of weights in the CRF's pairwise potentials (binary relations, differences of the region centroids)	Two-layer CRF
[22]	Models edge potentials, by considering the object's vertical location and height	Tree graphical model
[21]	Learning of weights for binary relations as a convex optimization problem	Non-maxima suppression (NMS) post-processing heuristic
[36]	Binary spatial constraints, using support and confidence as selection criteria	Fuzzy constraint satisfaction problem

III. VISUAL ANALYSIS

A. Segmentation and Visual Feature Extraction

In order to perform the initial region-concept association, the examined image has to be segmented to regions and suitable low-level descriptions have to be extracted for every resulting segment. In this work, a modified K-Means-with-connectivity-constraint pixel classification algorithm has been used for segmenting the image [39]. Output of this segmentation algorithm is a segmentation mask, where the created spatial regions $s_n,\ n \in [1, N]$, are likely to represent meaningful semantic objects.

Every generated image segment s_n is subsequently represented with the use of a visual feature vector v_n. Two different methods for estimating v_n are considered. Regarding the first one, the following MPEG-7 descriptors are extracted and concatenated to form the region feature vector: Scalable Color, Homogeneous Texture, Region Shape and Edge Histogram. This results in a 433-dimensional low-level feature vector. The second method is based on the Scale-Invariant Feature Transform (SIFT) [40]. In particular, a set of keypoints are initially estimated for every region s_n, using a point-of-interest detector as well as a pre-determined image grid, and a SIFT descriptor vector (with 128 elements) is extracted at each keypoint. Then, following the 'Bag-of-Words' (BoW) methodology [41], a 'vocabulary' of 300 visual words is constructed by performing clustering in the 128-dimensional feature space. Subsequently, each region is represented by the histogram of the visual words that it contains, i.e. the set of words that correspond to the original SIFT descriptors extracted from it. The latter histogram constitutes in this case the region feature vector v_n. The aforementioned visual features are in turn utilized by the classification algorithms, i.e. they constitute a common data set, for performing the region-concept assignment.

B. Visual Classification

In this section, the initial region-concept association procedure, i.e. the assignment of high-level semantic concepts to image regions based solely on visual information, is described. In the developed evaluation framework, three individual classification algorithms are employed: Support Vector Machines (SVMs), Random Forest (RF) and LogitBoost (LB). Every classifier receives as input either one of the two region feature vectors v_n described in Section III-A and estimates for every defined concept c_k, $k \in [1, K]$, a posterior probability $h_{nk} \equiv P(c_k|v_n)$. This probability denotes the degree with which concept c_k is assigned to region s_n.

SVMs have been widely used in semantic image analysis tasks due to their reported generalization ability and their suitability for handling high-dimensional data [42]. Under the proposed approach, an individual SVM is introduced for every defined semantic concept c_k to detect the corresponding instances, and is trained under the '*one-against-all*' approach. Each SVM receives as input the region feature vector v_n and estimates the posterior probability h_{nk} as follows: $h_{nk} = \frac{1}{1+e^{-\eta \cdot z_{nk}}}$, where z_{nk} is the distance of the particular input feature vector v_n from the corresponding SVM's separating hyperplane and η is a slope parameter set experimentally. This distance is positive in case of correct classification and negative otherwise.

RFs [43] belong to the general category of ensemble classifiers, i.e. classifiers that build on the combination of the outputs of multiple weak learners. In particular, the RFs' functionality is based on the combination of multiple decision tree classifiers, each of which is trained on different subsets of training samples and/or different subsets of features. RFs are considered to be robust to noisy data [43], while they are particularly suitable for data of high dimensionality or when a small number of training instances is present [44]. In this work, an individual RF classifier is defined for every supported concept c_k, while the '*one-against-all*' approach is followed for training. At the evaluation stage, the RF classifier estimates the posterior probability h_{nk} defined above, by averaging the outputs of the generated weak classifiers.

Boosting methods constitute a family of classification techniques that make decisions by combining the results of weak learners [45], similarly to the ensemble classifiers described above. The main advantage of these methods is that they have been shown to be less susceptible to overfitting occurrences than most learning algorithms. In the present analysis framework, a particular boosting algorithm is selected, namely the LB classifier, which makes use of a logit transform (log-odds ratio) for converting the weighted sum of the weak learners' output to a probability [45]. Similarly to the SVM

and the RF classification schemes, an individual LB classifier is constructed for every concept c_k, while the '*one-against-all*' approach is again followed for training. The posterior probability h_{nk} is estimated this time by making use of the aforementioned logit transform. Among the several different options that are available for selecting the weak learners, regression trees were used in this work.

IV. Spatial Context Acquisition

The first step in the application of any spatial context technique is the definition of an appropriate set of spatial relations. In the present analysis framework, fuzzy directional spatial relations are used to denote the order of objects in space. The set of supported directional relations, denoted by $R = \{r_\gamma,\ \gamma \in [1,\Gamma]\}$, comprises the following relations: Above, Right, Below, Left, Below-Right, Below-Left, Above-Right and Above-Left. These are estimated for every ordered pair of image regions $(s_n, s_m),\ n \neq m$, in parallel to visual feature extraction. Relation r_γ estimated for the region pair (s_n, s_m) is denoted by $r_\gamma(s_n, s_m) \in [0,1]$. A detailed description of their extraction procedure can be found in [46].

After the spatial relations extraction, a learning process is typically followed by each technique for spatial context acquisition. For this purpose, a set of manually annotated image content, denoted by D^1_{tr} and for which the fuzzy directional relations have been computed, is used. For every possible ordered pair of concepts (c_k, c_l) a corresponding set of relations, denoted by R^{c_k,c_l}, is formed. This set comprises all relations $r_\gamma(s_n, s_m),\ n \neq m$, that have been computed for all region pairs in D^1_{tr}, where concepts c_k and c_l have been manually assigned to regions s_n and s_m, respectively. Additionally, sets $R^{c_k,c_l}_\gamma \subset R^{c_k,c_l},\ \gamma \in [1,\Gamma]$, are also created, with respect to every individual spatial relation r_γ. Subsequently, each of the selected spatial context techniques applies its learning approach for spatial context acquisition. In particular, the BIP estimates a binary constraint for every concept pair (c_k, c_l) and every supported spatial relation r_γ, which is denoted by $T_\gamma(c_k, c_l)$. The latter is defined equal to 1 if concepts c_k and c_l are 'allowed' to be connected through relation r_γ, whereas it is set equal to 0 otherwise. Constraints $T_\gamma(c_k, c_l)$ are computed using support and confidence as selection criteria, and making use of sets R^{c_k,c_l}_γ. On the other hand, the EBM allows the use of fuzzy spatial constraints. These are utilized to denote the concepts 'expected' spatial arrangement and are calculated as follows:

$$\begin{aligned} \mathbf{r}_{n,m} &= [r_1(s_n,s_m),\ r_2(s_n,s_m)...r_\Gamma(s_n,s_m)]^T \\ \bar{\mathbf{r}}^{kl} &= [\bar{r}^{kl}_1,\ \bar{r}^{kl}_2...\bar{r}^{kl}_\Gamma]^T = E[\mathbf{r}_{n,m}],\ \forall(s_n,s_m) \in R^{c_k,c_l}, \end{aligned} \quad (1)$$

where $[.]^T$ denotes the transpose of a matrix and an individual mean vector $\bar{\mathbf{r}}^{kl}$ is calculated for every ordered concept pair (c_k, c_l). Moreover, the GA follows a more elaborate statistical learning approach that takes into account, apart from the mean values, the variance and the correlations between the spatial relations. This is achieved by the calculation of the covariance matrix $cov(\mathbf{r}^{kl})$ for every concept pair (c_k, c_l), according to the following equation:

$$cov(\mathbf{r}^{kl}) = E[(\mathbf{r}_{n,m} - \bar{\mathbf{r}}^{kl})(\mathbf{r}_{n,m} - \bar{\mathbf{r}}^{kl})^T],\ \forall(s_n,s_m) \in R^{c_k,c_l} \quad (2)$$

The estimation of the covariance matrix $cov(\mathbf{r}^{kl})$ results in a more complete representation of the concepts' spatial configuration than using the mean vector $\bar{\mathbf{r}}^{kl}$ alone.

Having acquired the appropriate spatial constraints, each technique aims at estimating an optimal region-concept association, i.e. assigning a final concept c_k to every region s_n, taking into account both visual and spatial information. This association of concept c_k with region s_n is denoted g_{nk}.

V. Spatial Context Techniques

A. Genetic Algorithm

GAs have been used in a wide variety of optimization problems, where they have been shown to outperform other traditional methods [47]. Under the proposed approach, a GA is employed for deciding on the optimal semantic image interpretation by treating image analysis as a global optimization problem, taking into account spatial contextual information. GAs constitute one of the most widely known global optimization methods [48], in the sense that in most cases they achieve to find the optimal solution (or a solution very close to the global optimum). The GA, being in principle a stochastic process, is not always guaranteed to converge to the global maximum, as no other stochastic optimization method is. However, through the tuning of the GA's parameters (like selecting a sufficiently large number of chromosomes in every population, choosing an appropriate selection operator, selecting a suitable crossover operator, adjusting the probabilities of mutation and crossover, etc.) the employed GA is adapted to the problem of spatial context exploitation and it is shown experimentally that it is capable of reaching a solution close to the optimal one (if not the global maximum). It must be noted that GAs are generally more robust in finding the globally optimal solution, compared to other common local search algorithms that iteratively shift among possible solutions and are thus more likely to converge to local maxima (e.g. gradient descend methods, quasi-Newton method, etc.) [48].

The developed GA employs an initial population of randomly generated chromosomes. Every chromosome Δ represents a possible solution, i.e. each gene assigns one of the defined concepts c_k to an image region s_n; therefore $\Delta = \{g_{nk},\ n \in [1,N]\}$. After the population initialization, new generations are iteratively produced by the application of evolutionary operators (selection, crossover and mutation) until the optimal solution is reached. The GA makes use of an appropriate fitness function for denoting the plausibility of every possible image interpretation, which has the form:

$$f(\Delta) = \frac{\sum_{n,m} V(g_{nk}, g_{ml})}{N(N-1)}, \quad (3)$$

where $V(g_{nk}, g_{ml}) \in [0,1]$ indicates the degree to which the g_{nk}, g_{ml} region to concept mappings are consistent with the acquired contextual and other (e.g. visual) information and $N(N-1)$ denotes the number of ordered region pairs that

are present in the examined image and which contribute to the summation in the numerator. Output of the GA is a final region-concept association which corresponds to the solution with the highest fitness value.

Regarding the acquisition of the appropriate spatial constraints, a statistical learning approach is followed. This involves the estimation of the set of values $\bar{\mathbf{r}}^{kl}$ and $cov(\mathbf{r}^{kl})$ (Section IV) for every possible concept pair (c_k, c_l), which represents the respective spatial constraint denoted u^{kl}. For evaluating the agreement of a given pair of region to concept mappings (g_{nk},g_{ml}) with spatial constraint u^{kl}, the following mahalanobis distance-based expression is used: $Y_{u^{kl}}(g_{nk}, g_{ml}) = \frac{1}{1+\sqrt{\mathbf{p}_{n,m}^T cov^{-1}(\mathbf{r}^{kl})\mathbf{p}_{n,m}}}$, where $\mathbf{p}_{n,m} = (\mathbf{r}_{n,m} - \bar{\mathbf{r}}^{kl})$. $Y_{u^{kl}}(g_{nk}, g_{ml}) \in [0, 1]$ denotes the degree to which the pair of mappings (g_{nk},g_{ml}) is consistent with the acquired spatial contextual information. Greater values of $Y_{u^{kl}}(g_{nk}, g_{ml})$ indicate more plausible spatial arrangements.

1) Combination of Spatial, Visual and Co-occurrence Information: The GA combines the available spatial with the visual and the concepts' co-occurrence information towards the detection of the most plausible pair of concepts (c_k, c_l) for each pair of regions. Concepts' co-occurrence indicates how often a given pair of concepts is observed. For performing this, i.e. estimating the value of $V(g_{nk}, g_{ml})$ in Eq. (3), a probabilistic approach is followed. In particular, a series of K^2 BNs are constructed, where an individual BN is introduced for every possible ordered pair of concepts (c_k, c_l) to learn the respective correlations. In the presented work, discrete space BNs are employed [49]. For every BN the following random variables are defined: a) variables CA_{nk} and CA_{ml}: CA_{nk} denotes the fact of assigning concept c_k to region s_n; similarly for CA_{ml}. b) variable SC_{nm}^{kl}, which represents the value of the spatial constraint verification factor $Y_{u^{kl}}(g_{nk}, g_{ml})$. c) variables VA_{nk} and VA_{ml}: VA_{nk} denotes the value of the estimated posterior probability h_{nk}; similarly for VA_{ml}. For variables CA_{nk} and CA_{ml} the set of values that they can receive is chosen equal to $\{ca_{nk1}, ca_{nk2}\} = \{ca_{ml1}, ca_{ml2}\} = \{True, False\}$. On the other hand, a discretization step is applied to the values $Y_{u^{kl}}(g_{nk}, g_{ml})$, h_{nk} and h_{ml} for defining the discrete values of random variables SC_{nm}^{kl}, VA_{nk} and VA_{ml}, respectively. The aim of the selected discretization procedure is to compute a close to uniform discrete distribution for each of the aforementioned variables. The structure of the BN defined for the concept pair (c_k, c_l) is denoted by $\mathbb{G}_{kl}$ and is illustrated in Fig. 2, where the direction of the arcs defines explicitly the causal relationships among the introduced random variables. From the developed BN structure $\mathbb{G}_{kl}$, the joint probability distribution of the random variables that it includes can be defined, according to the Markov condition [49]. This probability distribution is denoted by $P_{joint}(ca_{nk}, ca_{ml}, va_{nk}, va_{ml}, sc_{nm}^{kl})$, where ca_{nk}, ca_{ml}, va_{nk}, va_{ml}, sc_{nm}^{kl} are the values of the variables CA_{nk}, CA_{ml}, VA_{nk}, VA_{ml}, SC_{nm}^{kl}, respectively. It must be noted that the BN requires a set of annotated image content, denoted by D_{tr}^2 (similar to the D_{tr}^1 set described in Section IV), for training purposes.

At the evaluation stage, the BN receives as input the visual analysis results (i.e. posterior probabilities h_{nk} and h_{ml}) and the corresponding spatial constraint verification factor $Y_{u^{kl}}(g_{nk}, g_{ml})$. Then, it estimates the posterior probability $P(ca_{nk} = True, ca_{ml} = True | va_{nk}, va_{ml}, sc_{nm}^{kl})$, by performing inference. This probability constitutes a quantitative indication of how plausible the pair of region to concept mappings (g_{nk}, g_{ml}) is, based on spatial, visual and co-occurrence information; the value of $V(g_{nk}, g_{ml})$ in Eq. (3) is set equal to this probability. A detailed description of the overall spatial context technique can be found in [19].

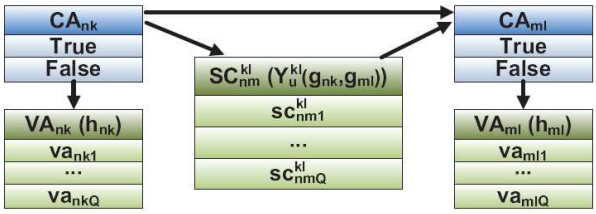

Fig. 2. Developed BN structure $\mathbb{G}_{kl}$ for combining spatial, visual and co-occurrence information. $va_{nk1}...va_{nkQ}$ denote discrete values of variable VA_{nk} and Q represents the total number of discrete values; similarly for VA_{ml} and SC_{nm}^{kl}.

B. Binary Integer Programming

Linear programs are a well known methodology for solving constraint satisfaction problems. BIPs are a specific type of linear programs, which allow the definition of only binary integer variables. Despite the complexity of BIPs being generally NP-hard, for certain forms they can be solved in polynomial time [50]. Under the proposed approach, the problem of spatial context exploitation is formalized as a BIP. For this purpose, a set of binary constraints $T_\gamma(c_k, c_l)$ (described in Section IV), which model the concepts' allowed spatial arrangement, need to be estimated. Then, the task of computing an optimal region-concept association is expressed in the form of a BIP, which can be solved efficiently and takes into account the initial classification results as well as the acquired spatial constraints.

For estimating the binary spatial constraints $T_\gamma(c_k, c_l)$, the set of spatial relations that can connect every concept c_k with any other concept c_l need to be determined. This is performed using support and confidence as selection criteria. For this purpose, additional sets of relations, apart from the sets $R_\gamma^{c_k,c_l}$ defined in Section IV, need to be generated from the image set D_{tr}^1. In particular, for every spatial relation r_γ a corresponding set of relations $R_\gamma^{c_k}$ is formed. This set comprises the relations $r_\gamma(s_n, s_m)$, $n \neq m$, that have been computed for all region pairs in D_{tr}^1, where concept c_k has been manually assigned to at least one of the regions s_n or s_m. Similarly, set R_γ^{*,c_l} is created, which contains all relations $r_\gamma(s_n, s_m)$ between any arbitrary region s_n and a region s_m associated with concept c_l. Then, the confidence value, denoted by $conf_\gamma(c_k, c_l)$, for spatial relation r_γ and concept pair (c_k, c_l) is calculated as follows: $conf_\gamma(c_k, c_l) = \frac{|R_\gamma^{c_k,c_l}|}{|R_\gamma^{*,c_l}|}$, where $|.|$ denotes the number of elements of a set. On the other hand, the corresponding

support value ($sup_\gamma(c_k, c_l)$) is estimated according to the following expression: $sup_\gamma(c_k, c_l) = \frac{|R_\gamma^{c_k,c_l}|}{|R_\gamma^{c_k}|}$. Spatial constraint $T_\gamma(c_k, c_l)$ is considered valid, i.e. $T_\gamma(c_k, c_l)$ is set equal to 1, if $conf_\gamma(c_k, c_l) > th_{conf}$ and $sup_\gamma(c_k, c_l) > th_{sup}$; otherwise, $T_\gamma(c_k, c_l)$ is set equal to 0. The values of the thresholds th_{conf} and th_{sup} are estimated following an optimization procedure, where image set D_{tr}^2 (Section V-A1) serves as a validation set.

In order to represent the problem of concern, i.e. spatial context exploitation, as a binary integer program, a set of linear constraints for each spatial relation need to be defined [50]. In particular, let O_n be the set of all outgoing relations for region s_n, i.e. $O_n = \{r_\gamma(s_n, s_m),\ \forall m \neq n\}$, and E_n the respective set of incoming relations, i.e. $E_n = \{r_\gamma(s_m, s_n),\ \forall m \neq n\}$. Then, for every supported spatial relation r_γ a corresponding binary integer variable $b_{n\gamma m}^{kl}$ is defined, which represents the region-concept mappings g_{nk}, g_{ml} with respect to relation $r_\gamma(s_n, s_m)$. $b_{n\gamma m}^{kl} = 1$ denotes that the mappings g_{nk}, g_{ml} are valid, while $b_{n\gamma m}^{kl} = 0$ that they are not. Since every binary variable $b_{n\gamma m}^{kl}$ represents the assignment of concept pair (c_k, c_l) to the pair of regions (s_n, s_m) with respect to relation $r_\gamma(s_n, s_m)$, and only a single concept can be eventually assigned to every region, this restriction has to be added as a set of linear constraints: $\sum_{c_k}\sum_{c_l} b_{n\gamma m}^{kl} = 1,\ \forall r_\gamma(s_n, s_m)$. These constraints ensure that there is only one pair of concepts assigned to a pair of regions with respect to every spatial relation. However, the aforementioned constraints do not assure that a unique concept is associated with every region in the final image interpretation, since a pair of binary variables for two spatial relations involving the same region might assign different concepts. In order to avoid that, additional constraints that 'link' the introduced variables need to be defined. This can be accomplished by linking pairs of relations. For the case of the outgoing relations, this is performed as follows: A reference relation $r_\gamma \in O_n$ is arbitrarily chosen and subsequently constraints regarding all $r_\zeta \in O_n$, $\zeta \neq \gamma$, are defined. Let $r_\gamma(s_n, s_m)$ and $r_\zeta(s_n, s_p)$ be the two relations to be linked. Then, the following constraints are defined: $\sum_{c_l} b_{n\gamma m}^{kl} - \sum_{c_l} b_{n\zeta p}^{kl} = 0,\ \forall c_k$. The first sum receives the value 1 if c_k is assigned to s_n with respect to relation r_γ. The second sum has to receive the same value, since both are subtracted and the whole expression has to be equal to 0. Therefore, if one of the relations assigns c_k to s_n, the other has to do the same. Following the same approach, the incoming relations, as well as the incoming with the outgoing ones, can also be linked. Eventually, an objective function, which denotes the plausibility of every possible image interpretation, is defined:

$$F = \sum_{r_\gamma(s_n,s_m)} \sum_k \sum_l \min(h_{nk}, h_{ml}) \cdot r_\gamma(s_n, s_m) \cdot T_\gamma(c_k, c_l) \cdot b_{n\gamma m}^{kl} \quad (4)$$

This function rewards concept assignments that satisfy the acquired spatial context and exhibit high analysis values (i.e. posterior probabilities h_{nk} and h_{ml}). The solution with the highest value of the objective function constitutes the output of the overall approach. A detailed description of this method can be found in [29].

C. Energy-based Model

EBMs are structured prediction models that encode the dependencies among the random variables that they include, while they can estimate an overall energy value for every possible combination of values of their random variables [38]. EBMs are defined in a way that more plausible sets of values of their random variables lead to lower energy levels. Inference aims at estimating the values of the defined random variables that minimize the overall energy of the model. EBMs are advantageous compared to other undirected graphical models that are widely used, like MRFs. This is mainly due to the fact that they allow the relaxation of the strict probabilistic assumptions and the avoidance of intractable partition functions that are often encountered in MRFs [38].

In this work, an improved version of the approach proposed in [37] is considered, which now incorporates information about the spatial arrangement of the image regions. The developed EBM reduces the region labeling problem to that of minimizing an energy function, which takes into account visual, spatial and co-occurrence information. In particular, the EBM is represented with a graph, where each node corresponds to a region s_n of the examined image. Dependencies among regions are denoted by edges. Under the proposed approach, all possible connections between the nodes of the model are considered and its general structure is illustrated in Fig. 3. Every node assigns one of the supported concepts c_k to every region s_n; this assignment is denoted g_{nk}, as described in Section IV. Additionally, the energy-function of the EBM for a given image is defined according to the following equations:

$$\begin{aligned} E &= -\Big(\sum_n t_1(g_{nk}) + \sum_{n,m} t_2(g_{nk}, g_{ml})\Big) \\ t_1(g_{nk}) &= \beta \cdot h_{nk} + \delta \cdot freq(c_k) \\ t_2(g_{nk}, g_{ml}) &= \mu \cdot freq(c_k, c_l) \cdot h_{ml} + \nu \cdot \phi(g_{nk}, g_{ml}) \end{aligned} \quad (5)$$

Term $t_1(g_{nk})$ in the above equations denotes the degree with which region s_n is associated with concept c_k, taking into account visual information (posterior probability h_{nk} defined in Section III-B) as well as the prior probability of occurrence of concept c_k, $freq(c_k)$. The latter is defined as the percentage (relative frequency) of the overall regions s_n that are present in the images of set D_{tr}^1 (Section IV) and constitute instances of concept c_k. Parameters β and δ adjust the degree to which h_{nk} and $freq(c_k)$ should affect the value of $t_1(g_{nk})$, respectively. On the other hand, term $t_2(g_{nk}, g_{ml})$ indicates the consistency of the g_{nk}, g_{ml} region to concept mappings, based on spatial ($\phi(g_{nk}, g_{ml})$) and co-occurrence ($freq(c_k, c_l)$) information. $freq(c_k, c_l)$ is defined equal to the percentage (relative frequency) of the region pairs (s_n, s_m) that are manually associated with the concepts (c_k, c_l) in the images of set D_{tr}^1. Additionally, factor $\phi(g_{nk}, g_{ml})$ is estimated using a normalized Euclidean distance-based formulation: $\phi(g_{nk}, g_{ml}) = 1 - \frac{||\bar{\mathbf{r}}^{kl} - \mathbf{r}_{n,m}||}{\sqrt{\Gamma}}$, where $||.||$ denotes the norm of a vector and the mean vector $\bar{\mathbf{r}}^{kl}$, which denotes

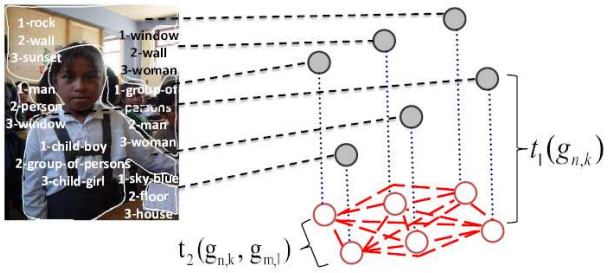

Fig. 3. Developed EBM for spatial context exploitation. Filled nodes denote the region-concept assignment based only on visual information, while unfilled ones represent the assignment after the EBM inference is performed.

the spatial constraint for concept pair (c_k, c_l), is calculated according to Eq. (1). The impact of factors $freq(c_k, c_l)$ and $\phi(g_{nk}, g_{ml})$ on the estimation of term $t_2(g_{nk}, g_{ml})$ is adjusted through parameters μ and ν, respectively. It must be noted that for selecting the optimal values for parameters β, δ, μ and ν, a grid search strategy is followed, where image set D^2_{tr} (Section V-A1) serves as a validation set.

At the evaluation stage, the EBM receives as input the visual analysis results (i.e. posterior probabilities h_{nk}), as well as the spatial relations $\mathbf{r}_{n,m}$ that have been computed for every possible region pair (s_n, s_m). Then, it assigns a particular concept c_k to every region s_n, ensuring that its overall energy value E (Eq. (5)) is minimized. In the current implementation, the Iterated Conditioned Modes (ICM) algorithm [51], i.e. an algorithm commonly used in EBM-based schemes, is utilized for realizing inference. The latter was experimentally shown to outperform other widely used methods like simulated annealing and graph cuts.

D. Discussion

Having described the selected spatial context techniques, an in-depth theoretical analysis and comparison about the core idea behind each method is presented in this section. In particular, detailed comments are given regarding the following key points in spatial context exploitation: a) the complexity of the utilized spatial contextual information, b) the combination of the spatial, visual and concepts' co-occurrence information, and c) the spatial constraints enforcement procedure.

Regarding the complexity of the utilized spatial contextual information, the BIP technique estimates a set of binary constraints ($T_\gamma(c_k, c_l)$), as described in Section V-B. These constraints define whether any two concepts can be connected through a given spatial relation or not. Inevitably, this choice results into relatively coarse representation of the objects' spatial configuration, since no information is included regarding the degree to which a spatial relation should be satisfied by a particular pair of concepts. On the other hand, the EBM incorporates a finer representation by estimating the concepts' expected spatial arrangement (values $\bar{\mathbf{r}}^{kl}$ in Eq. (1)). Although this enables the quantitative description of the objects' usual relative position and allows the use of fuzzy spatial constraints, this representation still does not include information about the variations that are observed in the spatial relations that connect the concepts. Moving to a more detailed description of the concepts' spatial topology, the GA estimates, apart from the values $\bar{\mathbf{r}}^{kl}$, the variances and the correlations between the spatial relations (covariance matrix $cov(\mathbf{r}^{kl})$ in Eq. (2)) for every pair of concepts. In this way, the GA achieves a more complete representation of the concepts' spatial configuration than the other techniques, since it encodes both the concepts' expected relative position as well as the extent of the variation from it.

For combining the spatial, the visual and the concepts' co-occurrence information, the selected spatial context techniques follow different fusion strategies. In particular, the BIP makes use of a product operator for combining the spatial with the visual information, as described in Eq. (4). This approach performs under the fundamental assumption that the visual features and the spatial relations for any pair of regions constitute statistically independent quantities, regardless of the semantic concepts that are present in these regions. Additionally, the concepts' co-occurrence information is taken into account only implicitly and without gradation with respect to the individual concepts, through the value of the binary spatial constraint $T_\gamma(c_k, c_l)$. On the contrary, the EBM drops the aforementioned statistical independence assumption and follows a weighted sum approach for performing the information fusion. Specifically, the EBM estimates a set of global weight factors (parameters β, δ, μ and ν in Eq. (5)) for adjusting the impact of the visual features against the concepts' prior probabilities and the co-occurrence information against the spatial context. The main drawback of this methodology is that the computed global weights are likely not to be appropriate for all concepts. On the other hand, the GA follows a more elaborate probabilistically-learning approach for efficiently performing information fusion, separately for every possible pair of concepts. In particular, the GA employs a set of BNs, which enable it to identify concepts whose detection could be boosted by the incorporation of the spatial information and subsequently to adjust the impact of every information source. This is carried out through the estimation of the probability distribution $P_{joint}(ca_{nk}, ca_{ml}, va_{nk}, va_{ml}, sc^{kl}_{nm})$ for the BN structure in Fig. 2.

The methodology followed for enforcing the acquired constraints also affects the efficiency of the spatial context exploitation procedure. In the present analysis framework, all spatial context techniques make use of machine learning methods for this purpose. In particular, the EBM performs a mapping of every image region to an individual node of a graphical model that represents the image. The main disadvantage of these models is that the maximum a posteriori (MAP) estimation is usually intractable, especially for models that contain many nodes, and the final solution can only be approximated. As a consequence, the inference algorithms utilized by the graphical models are in general locally optimal; hence, they can be easily misguided or substantially affected by the presence of noise in the data. Additionally, the EBM depends significantly on the concepts' prior probabilities, due to the initialization procedure of its nodes. On the other hand, the GA is a global optimization method, as discussed in Section V-A. As such, the GA is generally expected to

be less likely to converge to local maxima than the EBM and to be less affected by the presence of noise, especially when the examined image contains a relatively large number of regions. Moreover, the developed GA employs an initial set of candidate solutions, which are randomly distributed in the solution space and in this way render the method less dependent on the concepts' prior probabilities. Similarly to the GA, the BIP is also a global optimization method. Consequently, the BIP is also expected to be less affected by the presence of a relatively increased number of regions in the image than the EBM.

The experimental evaluation will show how the above algorithmic differences affect the performance of spatial context exploitation both in overall and for individual concepts.

VI. Experimental Evaluation

A. Datasets

In the developed evaluation framework, six datasets denoted D_1-D_6 of varying complexity are utilized. Each dataset was divided to three sub-sets, namely D_{tr}^1, D_{tr}^2 and D_{te}. The first one, D_{tr}^1, was used by the classification algorithms for training and by the spatial context techniques for acquiring spatial contextual information. D_{tr}^2 was utilized for optimizing the parameters of the spatial context techniques, while D_{te} was used for evaluation. The selected datasets are:

i) D_1 comprises 535 images depicting only coastal scenes. An appropriate set of 7 concepts c_k, which represent meaningful real-world objects that can be present in images of the formed set, was defined. Then, every image was manually annotated, i.e. after the segmentation algorithm described in Section III-A was applied, a single concept was associated with every resulting image region.

ii) The SCEF[1] dataset, which is denoted by D_2 and was introduced in [52]; D_2 (10 concepts) constitutes a broader dataset than D_1, including images that belong to different semantic categories.

iii) The LabelMe dataset [53], where the 16 most dominant concepts were considered (i.e. concepts with at least approximately 100 instances in the dataset). It must be noted that for this dataset (D_3) hand-made image annotation at region-level (i.e. the number and the boundaries of the regions in the image are also manually determined) was originally available. In order to generate ground-truth image annotations following the application of an automatic segmentation algorithm, a procedure similar to the 'figure-ground segmentation' approach proposed in [54] was followed. Specifically, every image was initially segmented using the algorithm of [39]. Then, every created image region s_n was assigned one of the supported concepts c_k if the percentage (%) of its area corresponding to concept c_k, based on the provided hand-made image annotation, exceeded a pre-defined threshold; otherwise, s_n was considered an 'unknown' region. The value of this threshold was experimentally set equal to 66%, while the respective value in the work of [54] was equal to 50%.

iv) D_4 comprises 648 images belonging to the personal collection domain. An appropriate set of 17 concepts was defined for it and manual image annotation at region-level was performed.

v) The PASCAL VOC2010[2] dataset (D_5) for the segmentation competition. The dataset (20 concepts) for this particular competition was selected, since hand-made pixel-level image annotations were available for it. In order to generate ground-truth image annotations using an automatic segmentation algorithm, a procedure similar to the case of the D_4 dataset was followed.

vi) Finally, the MSRC[3] v2 dataset was also used. For this dataset (D_6) 21 semantic concepts are supported and hand-made pixel-wise image annotations are provided. To this end, a procedure similar to the cases of the D_4 and D_5 datasets was performed again for generating region-level image annotations using an arbitrary segmentation algorithm.

The partitioning of every utilized dataset to the image sets D_{tr}^1, D_{tr}^2 and D_{te}, as well as the supported concepts for each dataset, are illustrated in Table III.

In order to examine the way that the supported concepts are distributed among the images of each dataset, the concepts' co-occurrence frequency $freq(c_k, c_l)$ (Section V-C) is calculated, taking into account this time all images of the respective dataset. The estimated values are depicted in Fig. 4. As can be seen from this figure, most concept pairs in D_1 exhibit relatively high co-occurrence frequency. This is due to the fact that the images of D_1 depict only coastal scenes. On the other hand, many frequency values $freq(c_k, c_l)$ are close or equal to zero in D_2. This is caused by the fact that the images of D_2 belong to different semantic categories; hence, some concept pairs are likely not to co-exist. Moreover, it can be seen that the co-occurrence matrices of Fig. 4 become more sparse for datasets D_3-D_6, as a result of the increased number of concepts that are supported for each of them. Especially for datasets D_3, D_5 and D_6, the concepts' co-occurrence frequencies are particularly low (and many of them equal to zero). This is mainly due to each image of the aforementioned datasets depicting very few different kinds of objects (usually no more than two or three), and to only specific concept pairs usually co-existing. Another important characteristic that differentiates datasets D_3, D_5 and D_6 from the remaining ones concerns the number of the regions that are present in the image and do not correspond to any one of the defined concepts, i.e. the image regions considered as 'unknown' above. The percentage of these regions to the total number of segments is approximately equal to 58%, 70% and 42% in D_3, D_5 and D_6, respectively. This is caused by the significantly large number of image pixels that were manually determined as 'void' during the original hand-made annotation

[1] http://mklab.iti.gr/project/scef

[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/index.html

[3] http://research.microsoft.com/en-us/projects/objectclassrecognition/

TABLE III
UTILIZED DATASETS

Dataset	Number of images			Supported concepts		
	D_{tr}^1	D_{tr}^2	D_{te}			
D_1	131	132	272	c_1: sand c_4: vegetation c_7: sky	c_2: sea c_5: rock	c_3: boat c_6: person
D_2	230	230	462	c_1: building c_4: person c_7: sand c_{10}: snow	c_2: foliage c_5: road c_8: sea	c_3: mountain c_6: sailing-boat c_9: sky
D_3	1183	1183	1100	c_1: building c_4: road c_7: grass c_{10}: ground c_{13}: wall c_{16}: person	c_2: sky c_5: plant c_8: sidewalk c_{11}: car c_{14}: door	c_3: tree c_6: window c_9: water c_{12}: mountain c_{15}: sea
D_4	162	162	324	c_1: building c_4: vegetation c_7: person c_{10}: tree c_{13}: sea c_{16}: gradin	c_2: roof c_5: dried-plant c_8: sky c_{11}: trunk c_{14}: road c_{17}: board	c_3: grass c_6: ground c_9: rock c_{12}: sand c_{15}: court
D_5	477	477	956	c_1: aeroplane c_4: boat c_7: car c_{10}: cow c_{13}: horse c_{16}: potted-plant c_{19}: train	c_2: bicycle c_5: bottle c_8: cat c_{11}: dining-table c_{14}: motorbike c_{17}: sheep c_{20}: tv-monitor	c_3: bird c_6: bus c_9: chair c_{12}: dog c_{15}: person c_{18}: sofa
D_6	148	147	296	c_1: building c_4: cow c_7: aeroplane c_{10}: car c_{13}: sign c_{16}: chair c_{19}: dog	c_2: grass c_5: sheep c_8: water c_{11}: bicycle c_{14}: bird c_{17}: road c_{20}: body	c_3: tree c_6: sky c_9: face c_{12}: flower c_{15}: book c_{18}: cat c_{21}: boat

of the images in these datasets. On the contrary, for datasets D_1, D_2 and D_4 the corresponding value is lower than 10%.

B. Effect of Image Segmentation

The segmentation performance is an important parameter in object-level context exploitation frameworks, while the segmentation results cannot be perfect. However, in order to ensure high quality of the computed segmentation masks and to reduce the influence of the segmentation error, the employed segmentation algorithm [39] was selected after conducting an empirical evaluation with other common techniques of the literature (e.g. normalized cuts [55], extensions [56], [57] of the Recursive Shortest Spanning Tree (RSST) algorithm [58], etc.). Additionally, the parameters of the algorithm of [39] were selected separately for every utilized dataset after experimentation, in order to accomplish high segmentation accuracy.

Regarding the selection of the appropriate segmentation level, an empirical evaluation with different sets of parameters for the employed segmentation algorithm was also performed. This resulted in various segmentations levels, ranging from coarse segmentation masks to over-segmented images. When very few segments were present in an image (e.g. up to three or four), the efficiency of the selected spatial context techniques was generally reduced. This was caused by the presence of multiple objects in a single region and the examination of very few spatial constraints. Additionally, the parameters that led to over-segmented images (e.g. more than twenty regions per image on average) resulted in reduced spatial context performance, too. This was mainly due to the significantly increased problem complexity and the division of a single object to multiple image segments in this case. The latter observation is consistent with the one described in [23], where the overall object recognition performance of the proposed spatial context technique was reduced when using over-segmented images. On the contrary, when: a) the total number of the generated regions was not very small or too big (e.g. around ten segments per image on average for most of the utilized datasets), and b) the remaining parameters of the employed segmentation algorithm were selected so as to compute accurate object localization, all the selected spatial context techniques were shown to introduce their highest concept detection performance improvement. Therefore, since segmentation efficiency affects all techniques in a similar way, a detailed quantitative evaluation study regarding the influence of image segmentation on the performance of the selected spatial context techniques was not included in the developed evaluation framework.

C. Analysis of Overall Concept Detection Results

In Table IV, quantitative performance measures from the application of the spatial context techniques are presented in terms of the overall concept classification accuracy for all possible combinations of low-level features and classification algorithms and all utilized datasets. Additionally, the difference in accuracy, which is calculated by subtracting the detection accuracy accomplished based only on visual features from the corresponding one obtained after using spatial context, is also given. The latter is depicted in parentheses. Accuracy is defined as the percentage of the image regions that are

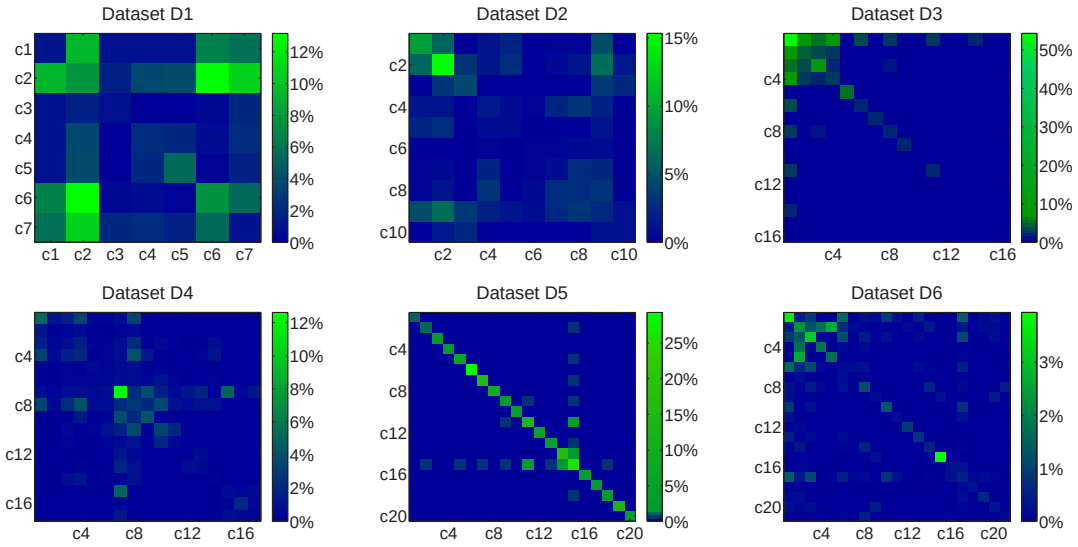

Fig. 4. Concepts' co-occurrence frequency in the utilized datasets.

associated with the correct semantic concept. In Figs. 5-10, the respective detailed concept detection results are also illustrated. It must be noted that for each region s_n, $argmax_k(h_{nk})$ is considered to indicate its concept assignment based solely on visual features.

From the results presented in Table IV, it can be seen that the application of all spatial context techniques leads to a significant improvement in the overall concept classification accuracy for most feature-classifier combinations in all datasets. The highest performance improvement is achieved by the BIP approach for the SIFT-SVM combination in D_4, where an increase of $9,25\%$ is observed. Additionally, the highest performance in absolute values is accomplished by the combination of SIFT features, RF classifier and the GA method for all datasets. The above results highlight the effectiveness of spatial context exploitation in improving the region-concept association results that have been generated based solely on visual information.

Another important remark concerns the performance improvement introduced by the spatial context techniques with respect to the initial concept classification results. In particular, for a given classifier in a given dataset, the low-level features resulting in better initial classification performance tend also to lead to greater performance improvement. The highest such difference is noticed for the BIP in D_5, where for the MPEG-7-SVM combination the introduced overall performance improvement is $0,17\%$, while for the SIFT-SVM combination the respective improvement is equal to $5,51\%$. Examining the above observation together with the results depicted in Figs. 5-10, it can be seen that the aforementioned difference in performance occurs when the initial classification results are good for most supported concepts and not only for a relatively small subset of them. Sound exception to this observation is noticed in D_1, where despite the significant difference in performance between the MPEG-7- and the SIFT-based classification results, no corresponding increase in the performance improvement introduced by the spatial context techniques is observed. For example, for the SIFT-RF combination, where the initial classification accuracy is equal to $80,60\%$, the corresponding performance improvement accomplished by the GA, BIP and EBM approaches is $2,55\%$, $0,52\%$ and $0,45\%$, respectively. The latter improvements are lower than the corresponding ones attained for the MPEG-7-RF combination. This suggests that when the initial classification accuracy exceeds an upper bound, which is indicated by the conducted experiments to be close to 80% for D_1, then the efficiency of spatial context techniques in introducing a performance improvement over these results is reduced. On the other hand, when not exceeding this upper bound, it can be seen that the highest initial classification performance (obtained for any possible feature-classifier combination) also leads every spatial context technique to its highest exhibited performance in all datasets. The only exceptions to this observation are the BIP in D_4, D_5 and EBM in D_5.

Comparing the performance of the presented spatial context techniques among the utilized datasets, it is shown that the overall concept detection improvement that they achieve over the initial classification results tends to increase when the corresponding number of supported concepts decreases. In particular, it can be seen from Table IV that for most cases the performance improvement increases concerning a particular technique and a given feature-classifier combination, when moving from dataset D_6 to D_1. This is mainly due to the following reasons: a) Considering the datasets from D_6 to D_1, the number of concepts reduces, which results in a corresponding reduction of the problem complexity. As a consequence, the selected spatial context exploitation approaches become less likely to be misguided when searching for the optimal image interpretation, which in turn facilitates them in efficiently discriminating between the defined concepts. b) Increase in the total number of supported concepts renders more likely many different concept pairs to present very similar spatial arrangements. For example, the concept pairs road-building and sand-sea in D_2 share very similar spatial configurations,

TABLE IV
OVERALL CONCEPT CLASSIFICATION ACCURACY

	D_1 dataset				D_2 dataset			
Features	Classifier		Spatial context		Classifier		Spatial context	
MPEG-7	SVM:	71,54%	GA: 79,25%	(**7,71**%)	SVM:	57,20%	GA: 63,68%	(**6,48**%)
			BIP: 76,85%	(5,31%)			BIP: 59,39%	(2,19%)
			EBM: 73,48%	(1,94%)			EBM: 58,58%	(1,38%)
	RF:	72,51%	GA: 76,03%	(**3,52**%)	RF:	59,36%	GA: 63,38%	(**4,02**%)
			BIP: 73,78%	(1,27%)			BIP: 59,29%	(-0,07%)
			EBM: 73,56%	(1,05%)			EBM: 60,51%	(1,15%)
	LB:	71,54%	GA: 75,43%	(**3,89**%)	LB:	58,01%	GA: 62,33%	(**4,32**%)
			BIP: 72,88%	(1,34%)			BIP: 59,76%	(1,75%)
			EBM: 72,66%	(1,12%)			EBM: 59,32%	(1,31%)
SIFT	SVM:	76,40%	GA: 83,00%	(**6,60**%)	SVM:	62,80%	GA: 70,74%	(**7,94**%)
			BIP: 80,00%	(3,60%)			BIP: 65,78%	(2,98%)
			EBM: 77,83%	(1,43%)			EBM: 65,57%	(2,77%)
	RF:	80,60%	GA: 83,15%	(**2,55**%)	RF:	66,76%	GA: 74,49%	(**7,73**%)
			BIP: 81,12%	(0,52%)			BIP: 70,07%	(3,31%)
			EBM: 81,05%	(0,45%)			EBM: 69,66%	(2,90%)
	LB:	78,13%	GA: 82,55%	(**4,42**%)	LB:	65,95%	GA: 73,58%	(**7,63**%)
			BIP: 80,15%	(2,02%)			BIP: 67,74%	(1,79%)
			EBM: 79,48%	(1,35%)			EBM: 68,28%	(2,33%)

	D_3 dataset				D_4 dataset			
Features	Classifier		Spatial context		Classifier		Spatial context	
MPEG-7	SVM:	51,73%	GA: 60,53%	(**8,80**%)	SVM:	50,81%	GA: 55,91%	(**5,10**%)
			BIP: 52,23%	(0,50%)			BIP: 53,39%	(2,58%)
			EBM: 53,07%	(1,34%)			EBM: 51,94%	(1,13%)
	RF:	55,61%	GA: 60,51%	(**4,92**%)	RF:	49,78%	GA: 54,78%	(**5,00**%)
			BIP: 55,31%	(-0,30%)			BIP: 49,57%	(-0,21%)
			EBM: 56,17%	(0,56%)			EBM: 51,34%	(1,56%)
	LB:	48,35%	GA: 53,37%	(**5,02**%)	LB:	47,53%	GA: 52,37%	(**4,84**%)
			BIP: 47,93%	(-0,42%)			BIP: 50,11%	(2,58%)
			EBM: 49,73%	(1,38%)			EBM: 50,38%	(2,85%)
SIFT	SVM:	60,13%	GA: 65,83%	(**5,70**%)	SVM:	57,31%	GA: 64,89%	(7,58%)
			BIP: 61,18%	(1,05%)			BIP: 66,56%	(**9,25**%)
			EBM: 60,67%	(0,54%)			EBM: 58,76%	(1,45%)
	RF:	64,74%	GA: 69,22%	(**4,48**%)	RF:	59,03%	GA: 67,53%	(**8,50**%)
			BIP: 65,13%	(0,39%)			BIP: 60,97%	(1,94%)
			EBM: 65,26%	(0,52%)			EBM: 61,18%	(2,15%)
	LB:	54,49%	GA: 59,03%	(**4,54**%)	LB:	57,10%	GA: 64,89%	(**7,79**%)
			BIP: 55,08%	(0,59%)			BIP: 61,24%	(4,14%)
			EBM: 55,61%	(1,12%)			EBM: 60,54%	(3,44%)

	D_5 dataset				D_6 dataset			
Features	Classifier		Spatial context		Classifier		Spatial context	
MPEG-7	SVM:	18,60%	GA: 27,82%	(**9,22**%)	SVM:	41,99%	GA: 47,88%	(**5,89**%)
			BIP: 18,77%	(0,17%)			BIP: 41,31%	(-0,68%)
			EBM: 21,07%	(2,47%)			EBM: 43,24%	(1,25%)
	RF:	17,81%	GA: 22,08%	(**4,27**%)	RF:	39,86%	GA: 45,37%	(**5,51**%)
			BIP: 18,26%	(0,45%)			BIP: 37,64%	(-2,22%)
			EBM: 18,85%	(1,04%)			EBM: 43,44%	(3,58%)
	LB:	9,72%	GA: 13,71%	(**3,99**%)	LB:	35,91%	GA: 41,41%	(**5,50**%)
			BIP: 6,32%	(-3,40%)			BIP: 34,07%	(-1,84%)
			EBM: 11,60%	(1,88%)			EBM: 38,13%	(2,22%)
SIFT	SVM:	28,49%	GA: 34,14%	(**5,65**%)	SVM:	42,57%	GA: 47,78%	(**5,21**%)
			BIP: 34,00%	(5,51%)			BIP: 45,46%	(2,89%)
			EBM: 30,57%	(2,08%)			EBM: 44,31%	(1,74%)
	RF:	29,33%	GA: 35,49%	(**6,16**%)	RF:	43,05%	GA: 48,46%	(**5,41**%)
			BIP: 31,24%	(1,91%)			BIP: 46,72%	(3,67%)
			EBM: 29,58%	(0,25%)			EBM: 47,01%	(3,96%)
	LB:	14,75%	GA: 20,82%	(**6,07**%)	LB:	36,58%	GA: 42,18%	(**5,60**%)
			BIP: 12,39%	(-2,36%)			BIP: 33,98%	(-2,60%)
			EBM: 16,35%	(1,60%)			EBM: 38,51%	(1,93%)

since the first concept usually corresponds to an image region that is bellow a segment that corresponds to the second concept in each pair. It must be noted at this point that the performance improvement obtained by each technique for most pairs of features-classifier is significantly higher in D_4 than the corresponding ones achieved in D_3, D_5 and D_6. This is observed despite the fact that the total number of supported concepts in all these datasets is comparable (i.e. 16, 17, 20 and 21 concepts are defined in D_3, D_4, D_5 and D_6, respectively). The latter is mostly caused by the following facts: a) Each image of datasets D_3, D_5 and D_6 depicts very few different kinds of objects and only particular concept pairs tend to co-exist, as discussed in Section VI-A. As a result, if regions are associated with an incorrect concept with a high posterior probability h_{nk} in images of these datasets, it is less likely to be eventually assigned the correct concept through the exploitation of spatial contextual information, compared to a similar case of images belonging to D_4. b) The percentage of 'unknown' regions in images of D_3, D_5 and D_6 is approximately more than four times the respective percentage in D_4, as described in Section VI-A. This type of regions contribute to the misleading of the inference procedure of all techniques; hence, limiting the effectiveness of spatial context in D_3, D_5 and D_6.

From the results presented in Table IV, it can be seen that the GA technique performs significantly better than the BIP and the EBM ones for most feature-classifier combinations in all datasets. The reason for this is twofold: a) The GA follows a more sophisticated statistical learning-based procedure for

acquiring complex fuzzy spatial constraints, compared to the simpler fuzzy constraints estimated by the EBM and the set of binary ones acquired by the BIP. b) The GA makes use of a BN-based approach for probabilistically adjusting the impact of the spatial, visual and concepts' co-occurrence information on the final outcome, separately for every pair of supported concepts. On the contrary, the EBM estimates a set of global weight factors and the BIP employs the product operator for performing the same task. The above observations indicate that acquiring complex spatial contextual information as well as efficiently adjusting its weight against the other information sources (e.g. visual, co-occurrence) can lead to a significant increase in the region-concept association performance. Moreover, these observations are also in accordance to the theoretical analysis given in Section V-D.

D. Discussion on Individual Concept-level Results

Having discussed the overall performance of each spatial context technique, their corresponding concept-level performance is examined here. The detailed concept detection results for all utilized datasets are given in Figs. 5-10, as described in Section VI-C. From the presented results, it can be seen that the selected spatial context techniques accomplish to significantly increase the detection rates for most of the supported concepts for any combination of low-level features and classifiers in all datasets. This fact demonstrates again the effectiveness of spatial context exploitation in improving the region-concept association results that have been computed based on visual features.

In order to evaluate the performance of every technique for each concept individually, a gradation of the supported concepts for each dataset is performed, with respect to how well-defined their spatial configuration is. Although there is no generally applicable formula for that purpose, the following quantitative measure, called Spatial Context Factor (SCF), is considered in this work:

$$SCF(c_k) \;=\; \frac{\sum_l tr(cov(\mathbf{r}^{kl})) + \sum_l tr(cov(\mathbf{r}^{lk}))}{2K}, \qquad (6)$$

where $tr(.)$ denotes the trace of a matrix and the covariance matrices $cov(\mathbf{r}^{kl})$ and $cov(\mathbf{r}^{lk})$ are calculated according to Eq. (2). Concept c_k is considered to have well-defined spatial context if the factor $SCF(c_k)$ receives relatively low values, i.e. the spatial relations of concept c_k with all other concepts c_l of the respective dataset do not present significant variations in their values. In Table V, the values of the factors $SCF(c_k)$, which are calculated for all supported concepts c_k in all utilized datasets, are presented in ascending order. Additionally, the weighted average value of the factors $SCF(c_k)$ is also given for each dataset. The latter constitutes a global quantitative indicator of the degree to which the supported concepts have well-defined spatial context for every dataset and is calculated according to the following expression: $\sum_k freq(c_k) \cdot SCF(c_k)$, where $SCF(c_k)$ and $freq(c_k)$ (Section V-C) are calculated, taking into account all images of the respective dataset. For example, concepts sand and sky exhibit relatively low values of factor $SCF(c_k)$ (i.e. they have more well-defined spatial context) in D_1, namely $0,3054$ and $0,3070$, respectively. This is due to the fact that sand instances in coastal scene types usually correspond to regions at the bottom of an image and are connected with relation below with most other concepts; similarly, sky instances are mainly connected with relation above with most other concepts. On the other hand, concept vegetation presents the highest $SCF(c_k)$ value, which is equal to $0,5139$. The reason for this is that vegetation may correspond to image regions with significantly different spatial configurations (e.g. tree foliage, bushes, etc.); hence, significant variations can be observed in their spatial relations with other concepts.

Examining the performance of the different spatial context techniques separately for every supported concept, it can be observed that the detection of some concepts is particularly favored by the application of each technique, regardless of the employed features-classifier combination. In particular, it is shown that concepts with more well-defined spatial context, according to the values of factors $SCF(c_k)$ depicted in Table V, exhibit the highest in percentage improvement over the initial classification results when the GA approach is applied. These concepts include: a) sand and person in D_1, b) sand and road in D_2, c) ground and mountain in D_3, d) court, board and grass in D_4, e) bus in D_5, and f) car and aeroplane in D_6. This suggests that the more sophisticated statistical learning approach followed for obtaining the GA's fuzzy spatial constraints u^{kl}, i.e. calculation of $\bar{\mathbf{r}}^{kl}$ and $cov(\mathbf{r}^{kl})$ (Section IV), is more suitable for modeling the spatial configuration of these concepts. Additionally, the EBM approach, which follows a simpler learning process than the GA for acquiring fuzzy spatial contextual information (i.e. only the $\bar{\mathbf{r}}^{kl}$ values are calculated), tends also to favor concepts with more well-defined spatial context, like concepts sky and road in D_6. On the other hand, the BIP technique, which makes use of a set of binary spatial constraints ($\mathcal{T}_\gamma(c_k, c_l)$), is shown to be advantageous for localizing concepts with not so well-defined spatial context, like sea, person and building in D_2, D_4 and D_6, respectively.

From the presented results, it can also be seen that significant performance improvement can be obtained for concepts that exhibit low initial classification rate by the application of the GA, like concepts: a) building, sailing-boat and snow in D_2, b) building in D_4, and c) bicycle, flower and chair in D_6. Significant contribution towards this performance improvement is induced by the probabilistic approach that is followed by the GA for adjusting the impact that the visual cues should have on the detection of every supported concept. On the contrary, marginal changes or decrease in the detection performance may be observed by the application of all techniques for concepts whose initial classification rate exceeds an upper bound similar to the one discussed in Section VI-C (e.g. concept sky in most datasets).

E. Effect of the Number of Image Regions and the Amount of Data Used for Context Acquisition

In this section, the performance of each technique is examined with respect to the number of regions that are present in the examined image, and with respect to the amount of

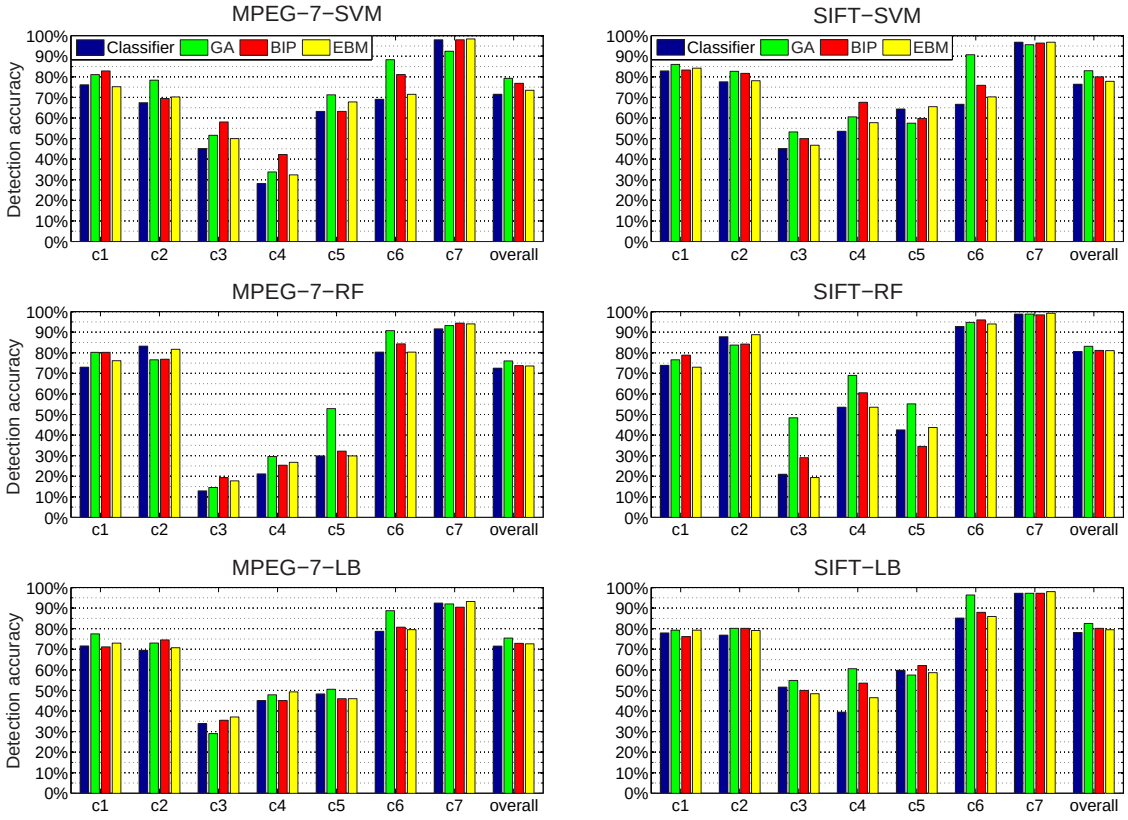

Fig. 5. Concept classification results in the D_1 dataset.

the available image content that is used for spatial context acquisition.

Regarding the first experiment, the images of test sets D_{te} for D_1 to D_6 are initially grouped with respect to the total number of regions they contain. Subsequently, the concept detection accuracy, as defined in Section VI-C, is calculated for each group of images and for every possible feature-classifier combination with respect to every spatial context technique. Then, for each technique a weighted average classification value is estimated for every corresponding group of images, using the number of images in each group of every dataset as weight. The obtained results are illustrated in Fig. 11(a) in terms of the difference in concept detection accuracy, i.e. by subtracting the average classification obtained based on visual features from the corresponding one accomplished by each spatial context technique. Additionally, the total number of images of all datasets in each of the aforementioned groups is given in Fig. 11(b). From the presented results, it can be seen that for the GA, which performs better on average than the other two methods in all datasets (as discussed in Section VI-C), a gradual increase in its performance improvement is observed, when the number of image regions increases. In particular, it is shown that when the number of regions N is equal to 2, the GA introduces a decrease of approximately $-1,58\%$ in the classification accuracy, compared to the initial classification results. On the other hand, significant performance improvement is observed when $N \geq 4$, exhibiting a highest value of approximately $14,46\%$ for $N = 27$. The reason for this is twofold: a) When very few regions are present in the image (e.g. $N = 2, 3$), the final region-concept associations are strongly dependent on the initial classification results, since very few spatial constraints can be taken into account. On the contrary, when more regions exist, spatial relations between significantly more region pairs are considered before reaching the final decision; hence, it is more likely for regions that have been misclassified based on visual information to be eventually associated with the correct concept. b) The GA is a global optimization method (Section V-A); therefore, it is less likely to converge to local maxima in the solution space and it can efficiently utilize the increased number of the available spatial constraints when N receives high values. The BIP, on the other hand, presents its highest average performance improvement when the number of regions is significantly high (i.e. $N \geq 10$), since it is also a global optimization method (Section V-B). However, the reason that it fails to introduce performance improvements comparable to those of the GA is mainly the limitation of the binary spatial constraints $T_\gamma(c_k, c_l)$ (Section IV) that the BIP makes use of to model the concepts' spatial arrangement, as discussed in Section VI-C. On the contrary, the EBM exhibits a relatively constant performance improvement with small variations around the value of $1,50\%$ for any N. This is mainly caused by: a) the EBM, as being a graphical model-based approach, depends heavily on the concepts' prior probabilities, and b) the ICM algorithm (Section V-C), which is used by the developed EBM for performing inference, is a local search method; hence, it

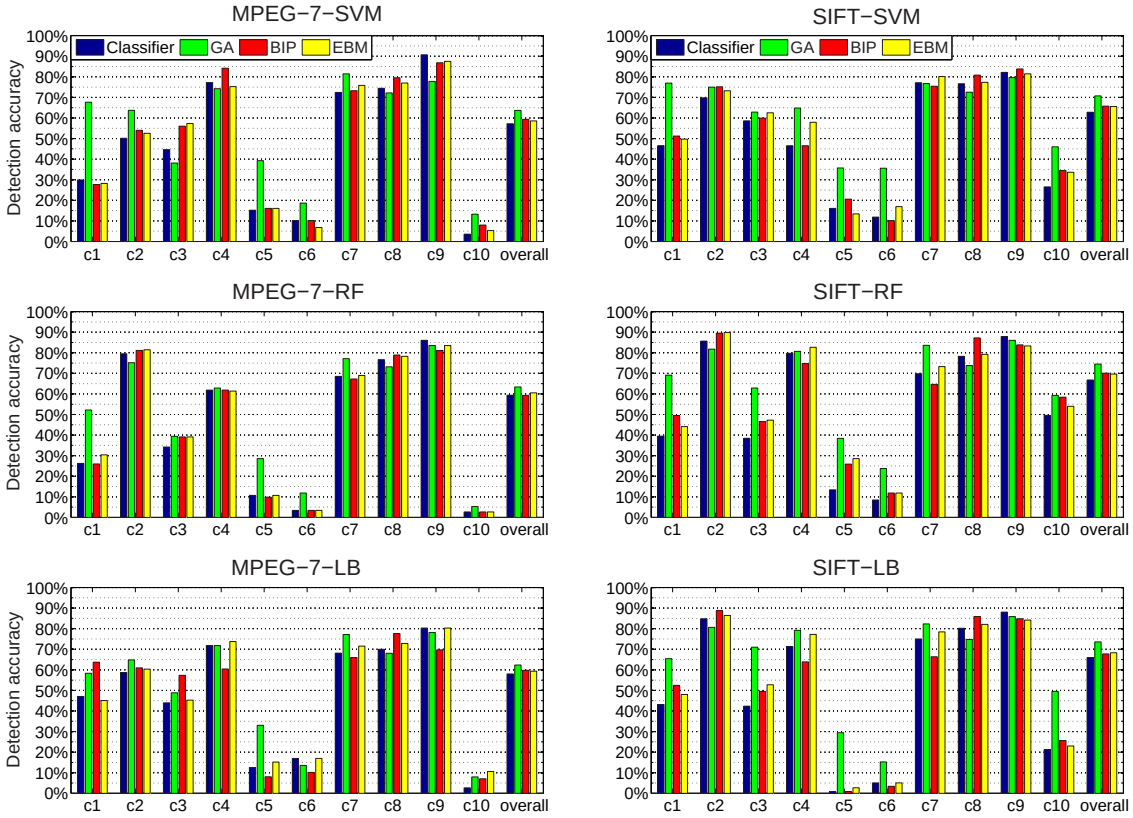

Fig. 6. Concept classification results in the D_2 dataset.

can not exploit efficiently the large number of constraints that are available in images with many regions, as opposed to the GA that follows a global optimization methodology. The above observations suggest that the adoption of a global optimization approach, while using more complex spatial contextual information, can lead to significant performance improvements over the initial classification results, when not very few regions exist in the examined image. These observation also justify the theoretical analysis given in Section V-D.

The performance of the spatial context techniques is also evaluated when the amount of the available image content (i.e. the number of images used) that is utilized for spatial context acquisition is reduced. For this purpose, the image set D_{tr}^1, which is used for spatial constraints learning (Section IV), is reduced to a corresponding $\acute{D}_{tr}^1$ one of half size, by randomly discarding half of its images. Then, the spatial context acquisition process of each technique is repeated, using $\acute{D}_{tr}^1$ instead of D_{tr}^1, and new region-concept association results are computed after the application of the GA, BIP and EBM techniques, as described in Section V. It must be noted that the initial classification results were maintained for ensuring a fair comparison. In Fig. 12, the obtained region-concept association results are given in terms of the difference in overall concept detection accuracy. The latter is calculated by subtracting the detection accuracy accomplished when using D_{tr}^1 for spatial context acquisition from the corresponding one obtained when $\acute{D}_{tr}^1$ is utilized. Additionally, the relation between the performance of the spatial context techniques and the initial classification based only on visual features is also illustrated (asterisks in Fig. 12). This is computed by subtracting the detection accuracy achieved when using D_{tr}^1 for spatial context acquisition from the detection accuracy obtained prior to the application of any spatial context technique. Bars higher than the respective asterisk in Fig. 12 indicate that the corresponding technique improves the initial classification results, after the reduction in the amount of image content used for spatial constraints learning. From the presented results, it can be seen that reducing the size of the set D_{tr}^1 to half results in small changes (i.e. changes $< 1\%$) in the performance of the GA for all possible feature-classifier combinations in all datasets. Additionally, its exhibited overall classification accuracy remains in all cases significantly higher than the baseline visual classification (i.e. the asterisks in Fig. 12). These observations indicate that the statistical learning approach followed by the GA for spatial constraints acquisition remains almost unaffected by the reduction in the amount of the available training data, despite the relatively more complex and sophisticated procedure that is followed, compared to the other two methods. The EBM follows the GA in the extent of the deviations in performance that it exhibits, with the highest reduction ($-3,19\%$) being observed in D_6. This is mainly due to the fact that the EBM supports the usage of fuzzy spatial constraints (similarly to the GA), following a simpler learning approach though, as described above. On the other hand, the BIP technique is shown to be affected the most by the reduction in the size of the available image content used

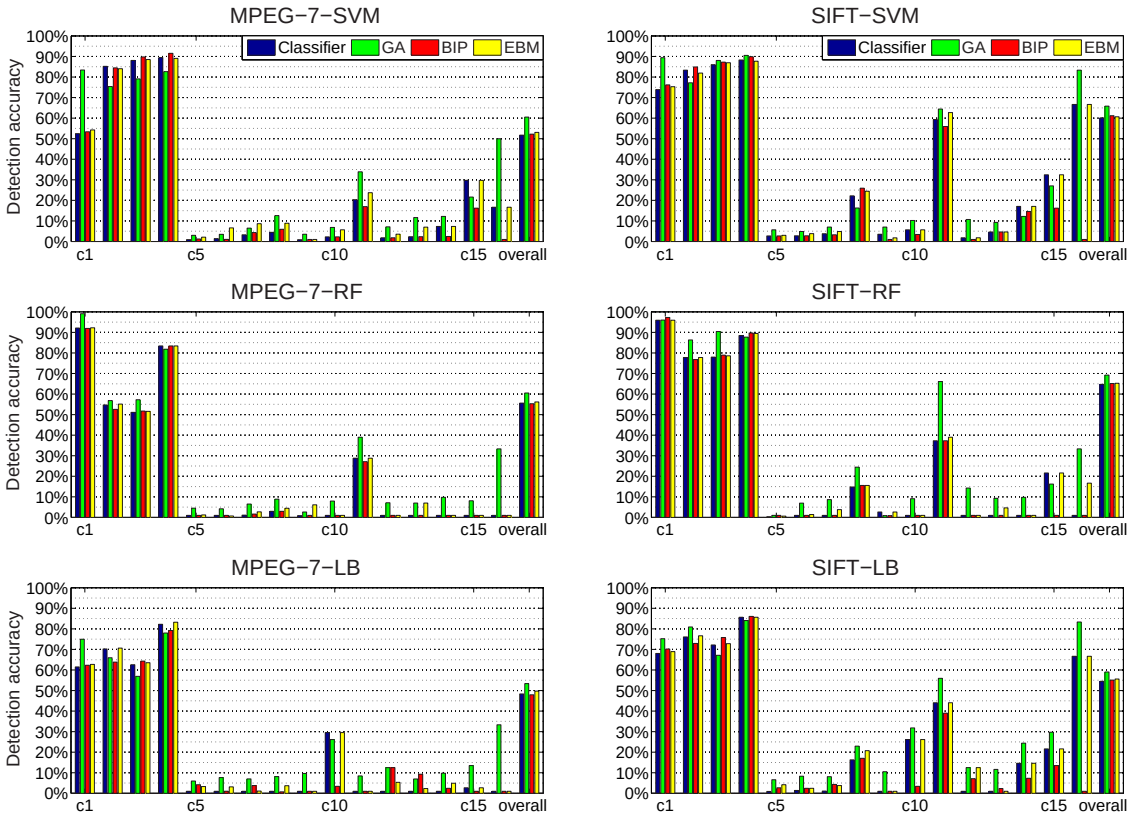

Fig. 7. Concept classification results in the D_3 dataset.

for spatial context acquisition. From the presented results, it can be seen that the highest reductions in performance are observed in D_4 and D_6, i.e. datasets with large number of supported concepts. In particular, a decrease in the overall performance equal to $-6,62\%$ and $-5,12\%$ is measured for the SIFT-LB combination in D_4 and D_6, respectively. These observations indicate that the set of binary spatial constraints used by the BIP, i.e. spatial relations that are either acceptable or not between two concepts, are more susceptible to reductions in the amount of training data, compared to the fuzzy constraints of the other two techniques.

F. Time Efficiency

In this section, the time efficiency of the selected spatial context techniques is investigated. In Table VI, the measured execution times of all techniques for both training and evaluation stages are given for all utilized datasets. The experiments[4] were conducted using a PC with Intel Quad Core processor at $2,4$GHz and a total of 3GB RAM.

From the results given in Table VI, it can be seen that the GA presents in general the highest execution times during the training stage and the BIP the lowest ones. The former is mainly due to the increased computational complexity of: a) the statistical learning approach that is followed by the GA for acquiring complex fuzzy spatial constraints, and b) the training procedure of the employed BNs for information fusion. On the contrary, the BIP appears to be the fastest method during training, since it adopts a significantly more simple approach for estimating a set of binary constraints. Regarding the time performance during the evaluation stage, it can be seen that the BIP is the slowest approach. This denotes that the deterministic methodology followed by the BIP for the constraints enforcement procedure is less efficient than the evolutionary approach of the GA and the local-search method of the EBM. Significant contribution to the latter have the binary constraints that the BIP technique makes use of, which have reduced expressiveness and which can hinder the efficient search of the optimal solution. On the other hand, the EBM is shown to perform the fastest. This is mainly due to the fact that the EBM uses a local search algorithm during its inference procedure, i.e. it does not follow a global optimization approach like the BIP and the GA techniques. It must be noted that the time efficiency of the GA, i.e. the most well-performing technique, depends heavily on the desirable level of the solution quality. The latter is mainly affected by the number of the chromosomes in the respective GA's population. In the current evaluation framework, more emphasis was given on reaching increased concept detection results than estimating an optimal trade-off between time efficiency and recognition performance for the GA technique (i.e. significant time improvements can be achieved with relatively low decrease in the concepts' recognition rates).

Regarding the time performance of each method among the

[4]Although the three spatial context techniques were executed with different PC configurations, appropriate time normalization based on hardware performance has been applied.

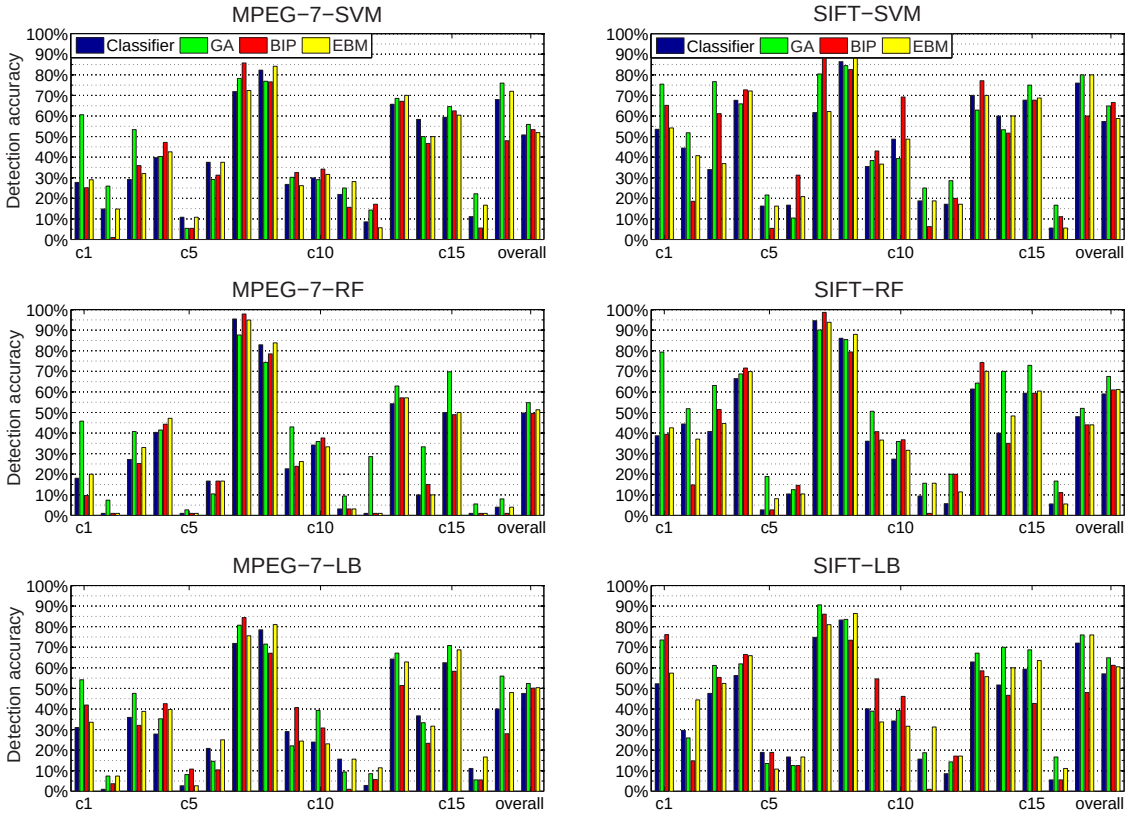

Fig. 8. Concept classification results in the D_4 dataset.

different datasets, it can be observed that the GA efficiency during the evaluation stage generally decreases as the number of images and the number of concepts increases. On the other hand, the BIP tends to present more intense and less consistent changes in time performance. Despite the inevitable randomness in the actual problem complexity in each dataset, this is mainly due to the significant variations in the number of the valid binary constraints that are learned for every dataset and which are likely to lead to significant time performance alterations. Additionally, it is shown that the time efficiency of the EBM mainly depends on the number of images in each dataset and it is not significantly affected by the number of the supported concepts. The latter is due to the local search strategy that it follows, which aims at estimating an approximation of the optimal solution when the problem complexity increases significantly. It must be highlighted though that the execution times between the datasets are not directly comparable, since the distributions of the total number of regions per image are different in every utilized dataset.

VII. Conclusions

In this paper, three approaches to spatial context exploitation that make use of fuzzy directional relations were presented and comparatively evaluated. The selected techniques include a GA, a BIP and an EBM, and each of them is applied after an initial set of region classification results based solely on visual features is computed. Extensive experiments on six datasets of varying complexity demonstrated the influence of a series of factors on their region-concept association performance. The main outcomes of this work regarding the exploitation of spatial context in semantic image analysis are summarized as follows:

- Spatial context is efficient in improving the initial (i.e. based solely on visual features) region-concept association results; exhibiting an overall increase of up to $9,25\%$ in the current evaluation framework.
- The highest on average performance is achieved when complex spatial constraints are acquired and their weight against the visual and co-occurrence information is efficiently adjusted (this is better accomplished by the BN-based approach followed by the GA, rather than the global weights of the EBM or the product operator of the BIP).
- The overall concept detection improvement over the initial classification results tends to increase when the number of supported concepts decreases.
- For a given classifier, the visual features that result in better initial classification performance also tend to lead to greater performance improvement when applying a spatial context technique.
- When the initial classification accuracy exceeds an upper bound (either overall, or at concept-level), the efficiency of spatial context in introducing a performance improvement over these results is reduced.
- For a given dataset, the highest initial classification performance leads also to the highest performance after the

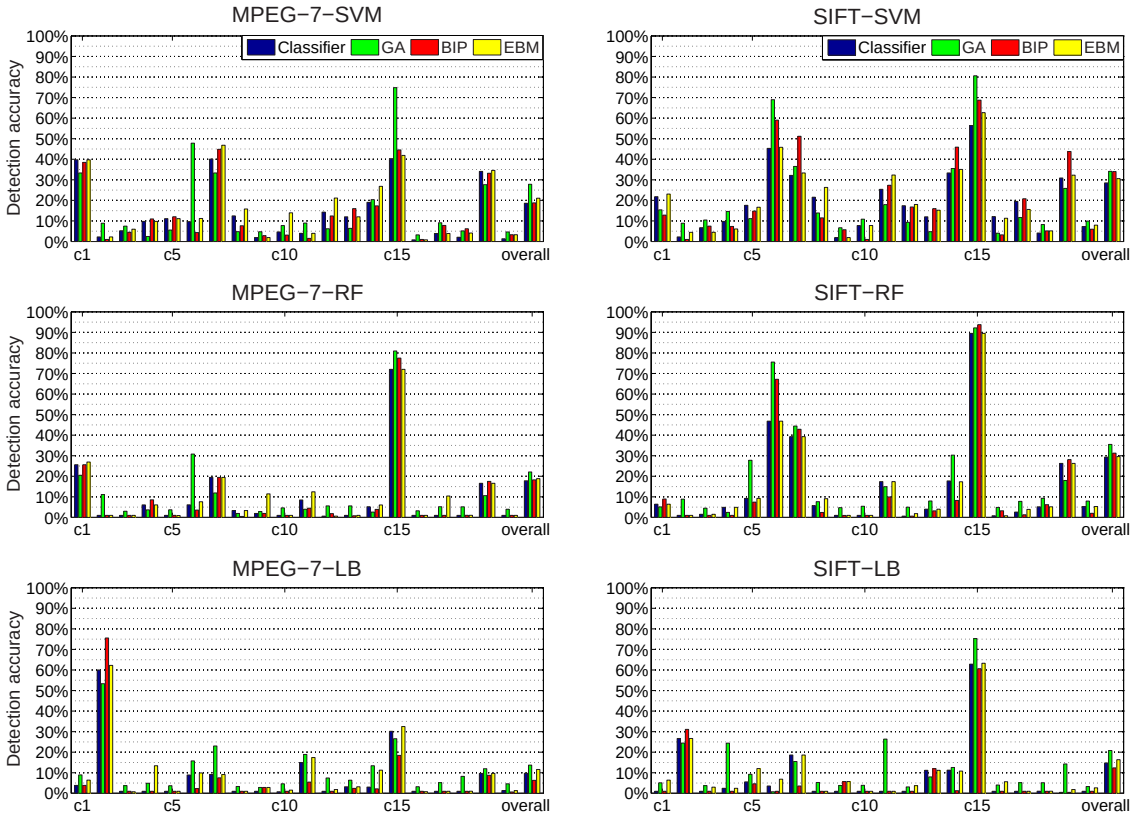

Fig. 9. Concept classification results in the D_5 dataset.

application of spatial context techniques.

- Fuzzy spatial constraints are less likely to result in performance decreases when the amount of training data is reduced, compared to binary constraints.

Additionally, the major differences in performance among the selected spatial context techniques, with respect to a series of individual factors, are given in Table VII. Future work includes the investigation of additional contextual information sources, like scene-type related information, and their combination with spatial context for achieving further performance improvement.

REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, pp. 1–60, 2008.

[2] A. Hanjalic, R. Lienhart, W. Ma, and J. Smith, "The holy grail of multimedia information retrieval: So close or yet so far away?" *Proceedings of the IEEE*, vol. 96, no. 4, pp. 541–547, 2008.

[3] S. Nikolopoulos, G. Papadopoulos, I. Kompatsiaris, and I. Patras, "An Evidence-Driven Probabilistic Inference Framework for Semantic Image Understanding," in *Machine Learning and Data Mining in Pattern Recognition, Proc. of the 6th Int. Conf. on*. Springer-Verlag, 2009, pp. 525–539.

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *Pattern Analysis and Machine Intelligence, IEEE Trans. on,*, vol. 24, no. 8, pp. 1026–1038, 2002.

[5] V. Mezaris, S. Gidaros, G. Papadopoulos, W. Kasper, J. Steffen, R. Ordelman, M. Huijbregts, F. de Jong, I. Kompatsiaris, and M. Strintzis, "A system for the semantic multi-modal analysis of news audio-visual content," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, p. 16, 2010.

[6] J. Luo, M. Boutell, and C. Brown, "Pictures are not taken in a vacuum," *IEEE Signal Processing Magazine*, vol. 23, 2006.

[7] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 712–722, 2010.

[8] A. Torralba, "Contextual priming for object detection," *Int. Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[9] J. Verbeek and B. Triggs, "Scene segmentation with conditional random fields learned from partially labeled images," in *Neural information processing systems (NIPS)*, 2008.

[10] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *IEEE Int. Conf. on Computer Vision*. IEEE, 2003, pp. 273–280.

[11] A. Singhal, J. Luo, and W. Zhu, "Probabilistic spatial context models for scene content understanding," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[12] G. T. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. Strintzis, "A Statistical Learning Approach to Spatial Context Exploitation for Semantic Image Analysis," in *Int. Conf. on Pattern Recognition, ICPR*, 2010.

[13] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *IEEE Int. Conf. on Computer Vision, ICCV*, vol. 2. IEEE, 2005, pp. 1284–1291.

[14] P. Carbonetto, N. Freitas, and K. Barnard, "A statistical model for general contextual object recognition," *Computer Vision-ECCV*, pp. 350–362, 2004.

[15] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. journal of computer vision*, vol. 81, no. 1, pp. 2–23, 2009.

[16] X. He, R. Zemel, and M. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," 2004.

[17] D. Parikh, C. Zitnick, and T. Chen, "From appearance to context-based recognition: Dense labeling in small images," in *Computer Vision and Pattern Recognition, IEEE Conf. on*, 2008, pp. 1 –8.

[18] L. Wolf and S. Bileschi, "A critical view of context," *International Journal of Computer Vision*, vol. 69, no. 2, pp. 251–261, 2006.

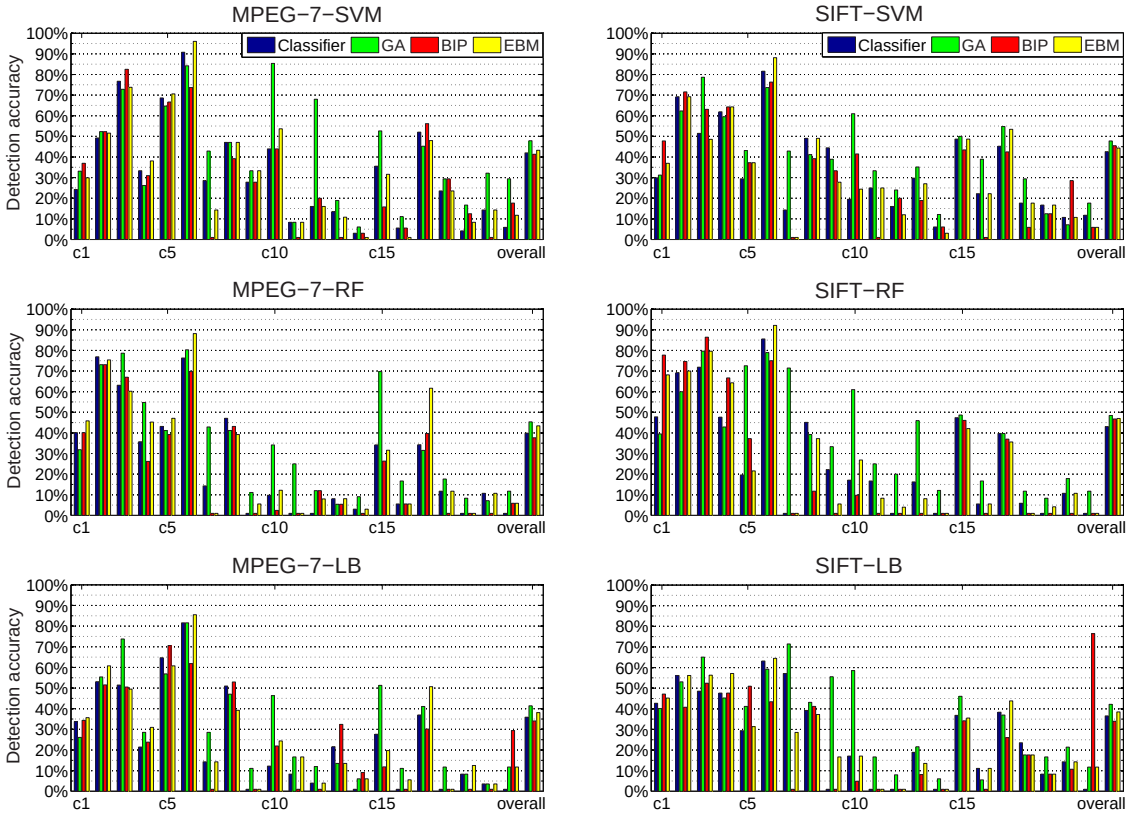

Fig. 10. Concept classification results in the D_6 dataset.

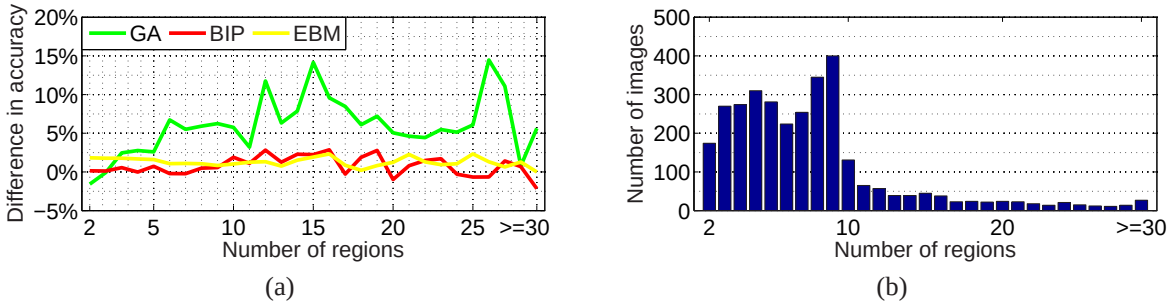

Fig. 11. (a) Average concept classification results and (b) number of images with respect to the number of regions they contain for all datasets.

[19] G. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. Strintzis, "Probabilistic Combination of Spatial Context with Visual and Co-occurrence Information for Semantic Image Analysis," in *IEEE Int. Conf. on Image Processing, ICIP*, 2010.

[20] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[21] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *Computer Vision, IEEE Int. Conf. on*. IEEE, 2009, pp. 229–236.

[22] M. Choi, J. Lim, A. Torralba, and A. Willsky, "Exploiting hierarchical context on a large database of object categories," *Computer Vision and Pattern Recognition, IEEE Computer Society Conf. on*, pp. 129–136, 2010.

[23] Z. Wang, D. Feng, Z. Chi, and T. Xia, "Annotating image regions using spatial context," in *Eighth IEEE Int. Symposium on Multimedia, ISM'06*, 2006, pp. 55–61.

[24] D. Semenovich and A. Sowmya, "A spectral method for context based disambiguation of image annotations," in *Proc. of IEEE Int. Conf. on Image processing*. IEEE Press, 2009, pp. 789–792.

[25] I. González-Dıaz, D. Garcıa-Garcıa, and F. de Marıa, "A spatially aware generative model for image classification, topic discovery and segmentation," in *IEEE Int. Conf. on Image Processing, ICIP*, 2009.

[26] J. Yuan, J. Li, and B. Zhang, "Exploiting spatial context constraints for automatic image region annotation," *Proc. ACM Int. Conf. on Multimedia*, pp. 595–604, 2007.

[27] M. Boutell, J. Luo, and C. Brown, "Improved Semantic Region Labeling Based on Scene Context," in *IEEE Int. Conf. on Multimedia and Expo*, 2005.

[28] D. Heesch and M. Petrou, "Markov Random Fields with Asymmetric Interactions for Modelling Spatial Context in Structured Scene Labelling," *Journal of Signal Processing Systems*, pp. 1–9, 2009.

[29] C. Saathoff, M. Grzegorzek, and S. Staab, "Labelling image regions using wavelet features and spatial prototypes," in *Int. Conf. on Semantics and Digital Media Technologies*, 2008.

[30] C. Hudelot, J. Atif, and I. Bloch, "Fuzzy spatial relation ontology for image interpretation," *Fuzzy Sets and Systems*, vol. 159, no. 15, pp. 1929–1951, 2008.

[31] A. Torralba, K. Murphy, and W. Freeman, "Contextual models for object detection using boosted random fields," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2005, pp. 1401–1408.

[32] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for

TABLE V
ESTIMATED $SCF(c_k)$ FACTORS

Dataset	Concepts			Average
D_1	sand (c_1): 0,3054	sky (c_7): 0,3070	sea (c_2): 0,4271	0,3978
	boat (c_3): 0,4505	person (c_6): 0,4509	rock (c_5): 0,4909	
	vegetation (c_4): 0,5139			
D_2	sand (c_7): 0,2831	sky (c_9): 0,3192	mountain (c_3): 0,4026	0,4255
	sea (c_8): 0,4150	road (c_5): 0,4261	sailing-boat (c_6): 0,4586	
	person (c_4): 0,4767	building (c_1): 0,5029	snow (c_{10}): 0,5070	
	foliage (c_2): 0,5188			
D_3	sea (c_{15}): 0,2492	water (c_9): 0,2880	sky (c_2): 0,3570	0,4409
	ground (c_{10}): 0,3571	grass (c_7): 0,4006	mountain (c_{12}): 0,4022	
	road (c_4): 0,4089	sidewalk (c_8): 0,4394	wall (c_{13}): 0,4485	
	building (c_1): 0,4543	plant (c_5): 0,4622	door (c_{14}): 0,4778	
	tree (c_3): 0,4923	person (c_{16}): 0,5092	car (c_{11}): 0,5113	
	window (c_6): 0,5234			
D_4	sand (c_{12}): 0,3081	road (c_{14}): 0,3319	court (c_{15}): 0,3454	0,4334
	sky (c_8): 0,3743	board (c_{17}): 0,3749	roof (c_2): 0,3825	
	sea (c_{13}): 0,3853	ground (c_6): 0,3922	rock (c_9): 0,4059	
	grass (c_3): 0,4060	gradin (c_{16}): 0,4187	building (c_1): 0,4638	
	dried-plant (c_5): 0,4669	vegetation (c_4): 0,4825	trunk (c_{11}): 0,4906	
	person (c_7): 0,5063	tree (c_{10}): 0,5378		
D_5	car (c_7): 0,3946	sheep (c_{17}): 0,4040	cat (c_8): 0,4290	0,4951
	boat (c_4): 0,4365	dog (c_{12}): 0,4572	aeroplane (c_1): 0,4580	
	bus (c_6): 0,4658	dining-table (c_{11}): 0,4668	motorbike (c_{14}): 0,4816	
	horse (c_{13}): 0,4858	bicycle (c_2): 0,4865	potted-plant (c_{16}): 0,5002	
	bottle (c_5): 0,5011	tv-monitor (c_{20}): 0,5118	sofa (c_{18}): 0,5343	
	bird (c_3): 0,5383	train (c_{19}): 0,5496	cow (c_{10}): 0,5537	
	chair (c_9): 0,5556	person (c_{15}): 0,5650		
D_6	sky (c_6): 0,2087	boat (c_{21}): 0,2482	sheep (c_5): 0,2675	0,3645
	water (c_8): 0,2747	cow (c_4): 0,3224	car (c_{10}): 0,3266	
	road (c_{17}): 0,3525	aeroplane (c_7): 0,3639	body (c_{20}): 0,3644	
	dog (c_{19}): 0,3737	cat (c_{18}): 0,3742	grass (c_2): 0,3776	
	bird (c_{14}): 0,3858	face (c_9): 0,3936	building (c_1): 0,4022	
	book (c_{15}): 0,4030	tree (c_3): 0,4218	chair (c_{16}): 0,4245	
	flower (c_{12}): 0,4576	sign (c_{13}): 0,4707	bicycle (c_{11}): 0,4912	

TABLE VI
TIME EFFICIENCY OF THE SELECTED SPATIAL CONTEXT TECHNIQUES IN MINUTES (TRAINING|EVALUATION)

Techniques	Datasets					
	D_1	D_2	D_3	D_4	D_5	D_6
GA	0,36|3,48	0,60|26,53	3,80|126,90	1,09|120,30	4,60|248,19	3,02|131,85
BIP	0,01|6,02	0,04|78,40	0,18|70,96	0,04|371,45	0,28|525,16	0,02|154,17
EBM	0,78|27,32	0,79|46,38	1,31|88,62	0,78|40,50	1,16|73,67	0,63|21,57

concurrent object segmentation and classification," in *Proc. ICCV*, 2007.

[33] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *IEEE Int. Conf. on Computer Vision, ICCV*, 2005, pp. 370–377.

[34] M. Fink and P. Perona, "Mutual boosting for contextual inference," in *Neural information processing systems*, vol. 1, 2004.

[35] A. Deruyver and Y. Hodé, "Qualitative spatial relationships for image interpretation by using a conceptual graph," *Image and Vision Computing*, vol. 27, no. 7, pp. 876–886, 2009.

[36] C. Saathoff and S. Staab, "Exploiting spatial context in image region labelling using fuzzy constraint reasoning," in *Image Analysis for Multimedia Interactive Services, Int. Workshop on*. IEEE, 2008, pp. 16–19.

[37] H. J. Escalante, M. Montes, and L. Sucar, "An Energy-based Model for Region-labeling," *Computer Vision and Image Understanding*, vol. 115, no. 6, pp. 787–803, 2011.

[38] G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan, *Predicting Structured Data*, ser. Advances in Neural Informacion Processing Systems. MIT Press, 2007, ch. Energy Based Models, pp. 191–246.

[39] V. Mezaris, I. Kompatsiaris, and M. Strintzis, "Still Image Segmentation Tools for Object-Based Multimedia Applications," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, pp. 701–726, 2004.

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Computer Vision*, vol. 60, pp. 91–110, 2004.

[41] Csurka, G. and Dance, C. and Fan, L. and Willamowski, J. and Bray, C., "Visual categorization with bags of keypoints," in *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, May 2004.

[42] V. Vapnik, *The nature of statistical learning theory*. Springer Verlag, 2000.

[43] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[44] T. Speed, *Statistical analysis of gene expression microarray data*. CRC Press, 2003.

[45] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[46] G. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. Strintzis, "Combining global and local information for knowledge-assisted image analysis and classification," *EURASIP Journal on Advances in Signal Processing, Special Issue on Knowledge-Assisted Media Analysis for Interactive Multimedia Applications*, vol. 2007, p. 15, 2007.

[47] M. Mitchell, *An Introduction to Genetic Algorithms*. MIT Press, 1996.

[48] P. Pardalos and H. Romeijn, *Handbook of global optimization. Volume 2*. Kluwer, 2002.

[49] R. Neapolitan, *Learning bayesian networks*. Prentice Hall Upper Saddle River, NJ, 2003.

[50] D. G. Luenberger, *Linear and Non-Linear Programming*, 1989.

[51] G. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*, ser. Applications of Mathematics. Springer, 2006, no. 27.

[52] G. Papadopoulos, C. Saathoff, M. Grzegorzek, V. Mezaris, I. Kompatsiaris, S. Staab, and M. Strintzis, "Comparative evaluation of spatial context techniques for semantic image analysis," *in Proc. WIAMIS '09*, pp. 161–164, 2009.

[53] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: a

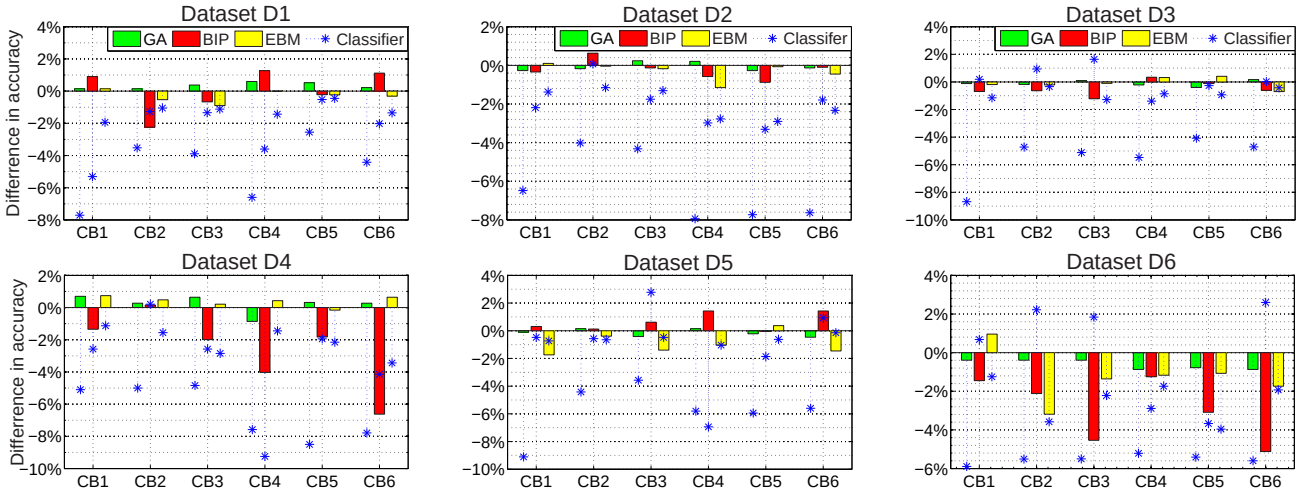

Fig. 12. Concept classification accuracy when using training set of reduced size (CB1: MPEG-7-SVM, CB2: MPEG-7-RF, CB3: MPEG-7-LB, CB4: SIFT-SVM, CB5: SIFT-RF, CB6: SIFT-LB). The asterisks represent the difference of the initial classification performance based only on visual features from the one accomplished when using D^1_{tr} for spatial context acquisition (i.e. bars higher than the respective asterisk indicate that the corresponding spatial context technique improves the results of the initial visual-based classification, after the reduction in the amount of image content used for spatial context acquisition).

TABLE VII
DIFFERENCES IN PERFORMANCE AMONG THE SPATIAL CONTEXT TECHNIQUES

Factors Considered	Spatial context techniques		
	GA	BIP	EBM
Concepts favored	Concepts with more well-defined spatial context and concepts with low initial classification rate	Concepts with less well-defined spatial context	Concepts with more well-defined spatial context
Number of image regions	Increase in performance improvement, when the number of regions increases ($N \geq 4$)	Performance improvement only when the number of regions is significantly high ($N \geq 10$)	Relatively constant performance improvement regardless of the number of regions
Reduction in amount of training data	Small changes in performance (changes $< 1\%$)	Significant performance reduction in datasets with many concepts can be observed (up to $-6,62\%$)	Performance reduction in datasets with many concepts can be observed (up to $-3,19\%$)

database and web-based tool for image annotation," *Int. Journal of Computer Vision*, vol. 77, no. 1, pp. 157–173, 2008.

[54] F. Ge, S. Wang, and T. Liu, "Image-segmentation evaluation from the perspective of salient object extraction," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2006, pp. 1146 – 1153.

[55] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 22, no. 8, pp. 888–905, 2000.

[56] T. Adamek, N. O'Connor, and N. Murphy, "Region-based segmentation of images using syntactic visual features," *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2005.

[57] T. Athanasiadis, P. Mylonas, Y. Avrithis, and S. Kollias, "Semantic image segmentation and object labeling," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 17, no. 3, pp. 298–312, 2007.

[58] O. Morris, M. Lee, and A. Constantinides, "Graph theory for image analysis: An approach based on the shortest spanning tree," *Communications, Radar and Signal Processing, IEE Proceedings F*, vol. 133, no. 2, pp. 146–152, 1986.