Evidence driven image interpretation by combining implicit and explicit knowledge in a bayesian network

Spiros Nikolopoulos, Georgios Th. Papadopoulos *Member, IEEE*, Ioannis Kompatsiaris *Member, IEEE* and Ioannis Patras *Member, IEEE*

***Abstract*—Computer vision techniques have made considerable progress in recognizing object categories by learning models that normally rely on a set of discriminative features. However, a drawback of those models is that, in contrast to human perception that makes extensive use of logic-based rules, they fail to benefit from knowledge that is provided explicitly. In this manuscript we propose a framework that is able to perform knowledge-assisted analysis of visual content. We use ontologies to model domain knowledge and a set of conditional probabilities to model the application context. Then, a bayesian network (BN) is used for integrating statistical and explicit knowledge and perform hypothesis testing using evidence-driven probabilistic inference. Additionally, we propose the use of a Focus of Attention (FoA) mechanism that is based on the mutual information between concepts. This mechanism selects the most prominent hypotheses to be verified/tested by the BN; hence, removing the need to exhaustively test all possible combinations of the hypotheses set. We experimentally evaluate our framework using content from three domains and for three tasks, namely image categorization, localized region labeling and weak annotation of video shot key-frames. The obtained results demonstrate the improvement in performance, compared to a set of baseline concept classifiers that are not aware of any context or domain knowledge. Finally, we also demonstrate the ability of the proposed FoA mechanism to significantly reduce the computational cost of visual inference, while obtaining results comparable to the exhaustive case.**

***Index Terms*—knowledge assisted image analysis, probabilistic inference, bayesian networks, ontologies, focus of attention.**

I. Introduction

THE advances in information technology have significantly reduced the traditional spatial and temporal obstacles in information exchange. Instant sharing infrastructures enable users to easily generate and exchange considerable amounts of digital data. However, the limitations of machine understanding makes it difficult for automated systems to interpret digital content in a manner coherent with human cognition, and the need for discovering intelligent ways to consume digital information is recognized as one of the emerging challenges of computer science [1]. With respect to multimedia, the difficulty of mapping a set of low-level visual features into semantic concepts, has motivated the use of domain knowledge for indexing this type of data. Moreover, as the importance of context in understanding audio-visual stimuli has been widely recognized, the integration of context and content is considered a promising approach towards multimedia understanding [2].

In our work we introduce a framework for enhancing image analysis using different types of evidence. Here, we define as evidence information that (when coupled with the principles of inference) can be used to support or disproof a hypothesis. In our framework (depicted in Fig. 1), we use visual stimulus, application context and domain knowledge to drive a probabilistic inference process that verifies or rejects a hypothesis made about the semantic content of an image. For a given task, the application context and the domain knowledge are considered to be the a priori/fixed information. On the contrary, the visual stimulus depends on the examined image and is considered to be the observed/dynamic information. In this manuscript, we propose a generative method for modeling the layer of evidence so as to effectively combine and exploit both a priori and observed information. More specifically, first we statistically analyze the visual stimulus to obtain conceptual information. Then, we represent domain knowledge and application context in a computationally enabled format. Finally, we combine everything in a bayesian network (BN) that is able to perform inference based on soft evidence. In this way, we provide the means to handle aspects like causality (between evidence and hypotheses), uncertainty (of the extracted evidence) and prior knowledge; hence, imitating some of human's basic perceptual operations when inspecting images.

The main contributions of our work can be summarized in the followings: a) We combine ontologies and bayesian networks for the purpose of allowing in a probabilistic way the fusion of evidence obtained at different levels of image analysis. We propose a data-oriented learning strategy for estimating the parameters of the BN. b) We show how global and regional evidence, as obtained from the application of concept classifiers on global and local image data respectively, can be probabilistically combined within a BN that incorporates domain knowledge and application context. We demonstrate that combining information in this way leads to statistically significant improvements for the tasks of image categorization, localized region labeling and weak annotation of video shot

Spiros Nikolopoulos is with CERTH/Informatics and Telematics Institute, Greece and with School of Electronic Engineering and Computer Science, Queen Mary University of London, UK (e-mail:nikolopo@iti.gr).

Georgios Th. Papadopoulos is with Electrical and Computer Engineering Department of Aristotle University of Thessaloniki, Greece and CERTH/Informatics and Telematics Institute, Greece (e-mail: papad@iti.gr)

Ioannis Kompatsiaris is with CERTH/Informatics and Telematics Institute, Greece (e-mail:ikom@iti.gr)

Ioannis Patras is with Electronic Engineering and Computer Science, Queen Mary University of London, UK (e-mail:i.patras@eecs.qmul.ac.uk)

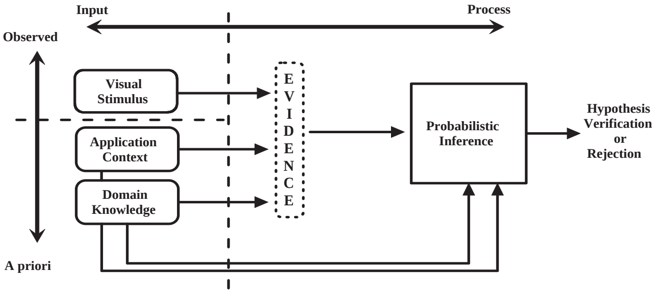


Fig. 1. Functional relations between the different components of the proposed framework.

key-frames. c) We propose a mechanism that exploits the mutual information between concepts, in order to significantly reduce the computational cost of visual inference and still achieve results comparable to the exhaustive case.

The rest of the manuscript is organized as follows. Section II reviews the related literature. Section III presents the individual components of the proposed framework. Section IV describes the methodology for migrating the semantic constraints expressed in an ontology into a BN. Section V details the functional settings of the proposed framework and Section VI describes our experimental study. Results are discussed in Section VII.

II. Related Work

Interpreting images in terms of their semantic content has been primarily addressed by devising methods that map low-level image visual characteristics (i.e., color, shape, texture) to high-level descriptions (i.e., semantic concepts), without making any use of domain knowledge and application context. Some indicative works that have been presented in the literature include [3] where the authors are based on scene-centered rather than object-centered primitives and use the mean of global image features to represent the gist of a scene, [4] where scene classification is performed using bayesian classifiers that operate on representations determined using a codebook of region types, and [5] where the authors introduce a visual shape alphabet representation with the aim to enable models for new categories to benefit from the detectors build previously for other categories. In this category of solutions we can also classify the methods that make combined use of global and local classification and treat images at a finer level of granularity, usually by taking advantage of image segmentation techniques. In [6] it is demonstrated through several applications how segmentation and object-based methods improve on pixel-based image analysis/classification methods, while in [7] a region-based binary tree representation incorporating adaptive processing of data structures is proposed to address the problem of image classification. Similarly, based on the combined use of local and global classification, [8] proposes a multi-level approach to annotate the semantics of natural scenes by using both the dominant image components (salient objects) and the relevant semantic objects, [9] employs Multiple-Instance-Learning to learn the correspondence between image regions and keywords and uses a bayesian framework for performing classification, while [10] presents a method where a new object is explained solely in terms of a small set of exemplar objects (represented as image regions). For each exemplar object a separate distance function is learned which captures the relative importance of shape, color, texture and position features. However, the inadequacy of the solutions relying solely on visual information to achieve efficient image interpretation has motivated the exploitation of context as a valuable source of information.

Context was defined in [11] as an extra source of information for both object detection and scene classification. Among the methods that make use of such information, we can identify the class of methods that develop models for spatial context-aware object detection, such as [12] that describes one generic outdoor-scene model, [13] that presents a model specific to individual archetypical scene types (e.g., beach, sunset, mountain, or urban), and [14] where multiple class object-based segmentation is achieved through the integration of mean-shift patches. Another class of methods that make use of such extra information includes the ones that exploit temporal context, as this can be derived from the surrounding images of an image collection (i.e., images drawn during a festival). In [15] the authors developed a general probabilistic temporal context model in which the first-order Markov property is used to integrate content-based and temporal context cues. Temporal context has been also used for active object recognition [16], as well as for identifying temporally related events [17]. Imaging context (i.e., camera metadata tags about scene capture properties, such as exposure time and subject distance) has been also used for aiding in a number of multimedia analysis tasks, including indoor-outdoor classification and event detection [18]. Other works that aim at improving the performance of individual detectors using contextual information are the ones that model the relationships between objects, such as [19] where contextual features are incorporated into a probabilistic framework which combines the outputs of several components, [20] where the authors present a two-layer hierarchical formulation to exploit the different levels of contextual information, and [21] where the authors propose a region-based model which combines appearance and scene geometry to automatically decompose a scene into semantically meaningful regions.

There is also a number of works that exploit conceptual context by developing techniques that are able to handle uncertainty and take advantage of domain knowledge. The authors of [22] introduce “Multijects” as a way to map time sequence of multi-modal, low-level features to higher level semantics using probabilistic rules. “Multinets” are also proposed for representing higher-level probabilistic dependencies between “Mutlijects”. In [23] “Multinets” are elaborated by introducing BNs for modeling the interaction between concepts and using this contextual information to perform semantic indexing of video content. A drawback of these approaches lies on the fact that the structure of “Multinets” is customly defined by experts and no methodology is suggested for explicitly incorporating the semantic constraints originating from the domain into the analysis process. In the same lines, [24] proposes a framework for semantic image understanding based on belief networks. The authors use three different image analysis tasks to demonstrate the improvement in performance introduced by extracting and integrating in the same knowledge-based inference framework (based on BNs), both low-level and semantic features. Once again, no systematic methodology is presented on how to seamlessly integrate domain knowledge, expressed with a standard knowledge representation language, into the probabilistic inference process. [25] describes an integrated approach of visual thesaurus analysis and visual context that exploits both conceptual and topological context. Another approach that attempts to model uncertainty and take advantage of knowledge and context for the task of multimedia analysis is [26]. This work uses low-level features and a BN to perform indoor versus outdoor scene categorization. In [27] a BN is utilized as an inference mechanism for facilitating a classification method based on feature space segmentation. Similarly, [28] propose a generative-model framework, namely dynamic tree-structure belief networks (DTSBNs), and formulates object detection and recognition as an inference process on a DTSBN. Domain knowledge is also used in [29], in order to tackle the problem that when training data is incomplete or sparse, learning parameters in BNs becomes extremely difficult. In their work the authors present a learning algorithm that incorporates domain knowledge into the learning process in order to regularize the otherwise ill-posed problem. Still, the absence of a methodology for integrating ontological knowledge into the inference process is what differentiates these works from our approach.

Works that utilize ontologies as a means to encode domain knowledge are also present in the literature. [30] presents a method for combining ontologies and BNs in an effort to introduce uncertainty in ontology reasoning and mapping. The Ontology Web Language (OWL) is augmented to allow additional probabilistic markups and a set of structural translation rules convert an OWL ontology into a directed acyclic graph of a BN. The conditional probability tables of the nodes are then calculated taking into consideration the ontology semantics. Probabilistic rules are used to cope with uncertainty and ontologies combined with belief networks are employed to express and migrate into a computationally enabled framework, the semantics originating from the domain. The proposed inference approach is validated using a synthetic example and no attempt is made to adjust the scheme for image analysis. [31] proposes a knowledge assisted image analysis scheme that combines local and global information for the task of image categorization and region labeling. In this case, a sophisticated decision mechanism that takes into account visual information, the concepts' frequency of appearance and their spatial relations is used to analyze images. [32] describes a scheme that is intended to enhance traditional image segmentation algorithms by incorporating semantic information. In this case, fuzzy theory and fuzzy algebra are used to handle uncertainty while a graph of concepts carrying degrees of relationship on its edges is employed to capture visual context. In [33] the authors build a concept ontology using both semantic and visual similarity in an effort to exploit the inter-concept correlations and to organize the image concepts hierarchically. In this process, the authors try to effectively tackle the problem of intra-concept visual diversity by using multiple kernels. However, none of [31], [32], [33] attempt to couple ontology-based approaches with probabilistic inference algorithms for combining concept detectors, context and knowledge. On the other hand, [34] uses ontologies as a structural prior for deciding on the structure of a BN, but in this work ontologies are mostly treated as hierarchies that do not incorporate any explicitly provided semantic constraints.

Finally, we should note that none of these works is concerned with computational efficiency and the fact that in a real-world inference system the number of plausible hypotheses could suffer from a combinatorial explosion. In this manuscript we discuss how visual inference can benefit from the use of exclusion principles and propose a focus of attention mechanism that is based on the mutual information between concepts.

III. Framework Description

A. Visual Stimulus

For analyzing the visual stimulus we consider the supervised learning paradigm where a classifier is trained to identify an object category, provided that a sufficiently large number of examples are available. We denote by N_C the set of domain concepts and by I_q the analyzed visual representation. Depending on the circumstances, I_q can be an image region, the whole image, a video shot, etc. A concept detector can then be implemented using a classifier F_c that is trained to recognize instances of the concept $c \in N_C$. We denote by $F_c(I_q)$ the output of F_c applied to image I_q. When F_c is a probabilistic classifier we have $F_c(I_q) = Pr(c|I_q)$. These probabilities $Pr(c|I_q)$ are essentially the soft evidence that are provided to the BN for triggering probabilistic inference.

B. Domain Knowledge

Let R be the set of binary predicates that are used to denote relations between concepts and O the algebra defining the allowable operators. In our framework we use the ontology web language OWL–DL [35] to construct a structure $K_D = S(N_C, R, O)$ that describes how the domain concepts are related to each other using R and $O \in DL$. DL stands for "Description Logics" [36] and constitutes a specific set of constructors such as intersection, union, disjoint, complement, etc. For instance, such constructors can be used to express that two concepts are disjoint with each other and can not be depicted in the same image simultaneously. Our goal is to use these constructors for explicitly imposing semantic constraints in the process of image interpretation that can not be captured by typical machine learning techniques. Loosely speaking, we use the knowledge structure to obtain: a) which of the domain concepts should be considered as evidence and therefore used to trigger the probabilistic inference process, and b) which evidence supports a certain hypothesis and what are the semantic restrictions that apply in this domain. In this sense, the knowledge structure sets the tracks to which evidence belief is allowed to propagate by determining the structure of the BN.

The use of ontologies instead of some other knowledge representation structure (e.g., conceptual graphs) was advocated by their wide acceptance and appeal in the area of knowledge engineering [37]. It is true that ontologies have been widely established as the main tool for encoding explicit knowledge in machine understandable format. This is witnessed by the fact that in many domains considerable effort has been already allocated on engineering ontologies that encode the existing concepts and relations. Therefore, enabling our framework to automatically handle ontologies makes it directly applicable in these domains.

C. Application Context

The role of K_D is to capture information about the domain in general, but not to deliver information concerning the context of the analysis process at hand. No information is provided to the framework regarding where within the content the anticipated evidence are likely to reside. For instance, this type of information could suggest the analysis mechanism to search for evidence in specific image regions. Moreover, information on how to quantitatively evaluate the existence of the extracted evidence (i.e., how much each hypothesis is affected by the existence of one evidence or another) is also missing from K_D. Let app denote the type of application specific information used to guide the analysis mechanism in searching for evidence, and $W = [W_{i,j}]$ the matrix whose elements $W_{i,j}$ quantifies the effect of concept c_i on c_j. Then, we consider the application context $X = S(app, W)$ to be the information consisting of both app and W. As will become clear in Section IV-B, W_{ij} is approximated by the frequency of co-occurrence between concepts c_i and c_j in the training set. This information, that is implicitly extracted from the training data, is encoded into the Conditional Probability Tables (CPTs) of the BN nodes and influences the probabilistic inference process when belief propagation takes place.

D. Evidence-driven Probabilistic Inference

In order to accommodate for evidence-driven probabilistic inference, our framework uses a BN derived from the domain ontology. This is accomplished by performing the following steps: a) we use K_D to decide which of the domain concepts should constitute the evidence set c^E, b) we use app to decide where to physically search for these evidence, c) we apply the probabilistic classifiers F_c on I_q to obtain the degrees of confidence for the concepts in c^E, d) we use app and K_D to decide which of the domain concepts should constitute the hypotheses set c^H, e) we provide the degrees of confidence for the concepts in c^E to the BN and trigger probabilistic inference by using these degrees as soft evidence, f) we propagate evidence beliefs using the network's inference tracks R and the corresponding causality quantification functions W_{ij}, and g) we calculate the posterior probabilities for all concepts in c^H and decide which of the hypotheses should be verified or rejected.

TABLE I
LEGEND OF INTRODUCED TERMS

Term	**Symbol**	**Role**
Trained Classifier	F_c	- Estimates the degree of confidence that the visual representation I_q depicts concept c.
Domain Knowledge	K_D=$S(N_C, R, O)$	- Determines which concepts belong to the evidence set and which to the hypothesis set. - Specifies qualitative relations between evidence and hypotheses (i.e., which evidence support a certain hypothesis)
Application Context	$X = S(app, W)$	- Determines where to "physically" search for evidence, expressed with app (i.e., application specific information). - Specifies quantitative relations (causality) between evidence and hypotheses, expressed with W.
Hypotheses	$h(I_q, c_i)$=$Pr(c_i\|I_q)$ and $H(I_q) = \{h(I_q, c_i) : c_i \in c^H\}$	- Constitutes the initial degrees of confidence for the concepts belonging to the hypotheses set c^H (as determined by $N_C \in K_D$ and $app \in X$) obtained by applying the classifiers F_c to I_q.
Evidence	$E(I_q) = \{h(I_q, c_i) : c_i \in c^E\}$	- Constitutes the degrees of confidence for the concepts belonging to the evidence set c^E (as determined by $N_C \in K_D$ and $app \in X$) obtained by applying the classifiers F_c to I_q.
Evidence driven Probabilistic Inference	$\acute{h}(I_q, c_i)$=$Pr(c_i \mid H(I_q), E(I_q), R, O, W_{ij})$ and $\acute{H}(I_q) = \{\acute{h}(I_q, c_i) : c_i \in c^H\}$	- Performs inference using $\acute{h}(I_q, c_i)$ and estimates the posterior probabilities $\acute{H}(I_q)$) using $E(I_q)$ as trigger, $R, O \in K_D$ as belief propagation tracks and $W \in X$ as causality quantification functions.
Semantic Image Interpretation	$c = \arg \otimes_{c_i \in c^H} (\acute{h}(I_q, c_i))$	Achieves semantic image interpretation based on the operator $\otimes$ that depends on the analysis task.

Let $h(I_q, c_i) = Pr(c_i|I_q)$ denote the function estimating the degree of confidence that concept c_i appears in image I_q. Let also $H(I_q) = \{h(I_q, c_i) : c_i \in c^H\}$ denote the set of confidence degrees that the concepts belonging to the hypotheses set are depicted in image I_q and $E(I_q) = \{h(I_q, c_i) : c_i \in c^E\}$ the set of confidence degrees that the concepts belonging to the evidence set are depicted in image I_q. Then, we provide $H(I_q)$ and $E(I_q)$ to the BN and using probabilistic inference we calculate the posterior probabilities of the network nodes using information coming from knowledge R, O and context W_{ij}. If we denote by $\acute{h}(I_q, c_i) = Pr(c_i \mid H(I_q), E(I_q), R, O, W_{ij})$ the function that calculates the posterior probabilities of the network nodes, the set of posterior probabilities of the concepts belonging to the hypotheses set can be represented as $\acute{H}(I_q) = \{\acute{h}(I_q, c_i) : c_i \in c^H\}$. The formula used to achieve semantic image interpretation can be expressed as follows:

$$c = \arg_{c_i \in c^H} \otimes (\acute{h}(I_q, c_i)) \tag{1}$$

$\otimes$ is an operator (e.g., $\max$) that depends on the specifications of the analysis task (Section VI describes the functionality of this operator for each of the analysis tasks). Table I shows the basic terms introduced in the proposed framework.

E. Computational Efficiency

Our evidence-driven probabilistic inference framework is essentially a method that connects a symbol (visual stimulus in our case) to real-world objects/concepts to which the symbol is associated. However, in the real-world the number of plausible hypotheses could suffer from a combinatorial explosion, rendering testing for them intractable. This problem is usually addressed using exclusion principles determined by the faculties of attention and perception [38]. In our case, the exclusion principles are derived from the domain ontology which determines the set of plausible hypotheses for each task.

Still, the computational cost for gathering the necessary evidence is often so expensive that can be prohibitive in highly complex domains. For this purpose we introduce a Focus of Attention (FoA) mechanism that improves the computational efficiency of the proposed framework. In particular, we apply an iterative process that initially examines the hypothesis and evidence that are more likely, in statistical terms, to be valid. If the hypothesis is verified the process is terminated, otherwise the next most likely hypothesis is examined. More specifically, instead of examining the complete hypotheses set $H(I_q) = \{h(I_q, c_i) : c_i \in c^H\}$, we initially examine the hypothesis with the maximum confidence degree c_k, where $k = arg\max_i(h(I_q, c_i))$ and $c_i \in c^H$. This is performed by inserting this value to the corresponding network node and comparing the node's posterior probability against a predefined belief threshold. If the posterior probability exceeds the belief threshold the process is terminated. Otherwise, a ranked list of the evidence concepts (i.e., $\forall c_i \in c^E$), that would have caused maximum impact on the hypothesis if were observed, is formed. This is performed by calculating the mutual information between the node corresponding to the concept c_k and all other nodes corresponding to the concepts of c^E. The mutual information between two discrete random variables is the expected reduction in entropy of one variable (measured in bits) due to a finding in the other variable. The mutual information between c_k and c_i, $\forall c_i \in c^E$ is calculated according the following equation:

$$I(c_k; c_i) = \sum_{\{true,false\}} \sum_{\{true,false\}} Pr(c_k, c_i) \log_2 \frac{Pr(c_k, c_i)}{Pr(c_k)Pr(c_i)}, \quad (2)$$

where $Pr(c_k, c_i)$ is the joint and $Pr(c_k)$, $Pr(c_i)$ the marginal probability distributions of c_k and c_i. The efficient calculation of $P_r(c_k, c_i)$ is performed using the junction tree [39], which is an efficient and scalable belief propagation algorithm that exploits a range of local representations for the network. Subsequently, the nodes are ranked in descending order based on their mutual information with c_k and the confidence degrees of the concepts corresponding to the most highly ranked nodes are extracted. The resulting degrees are inserted into the BN causing belief propagation to take place. If the posterior probability of the examined hypothesis still fails to exceed the pre-defined belief threshold, the hypothesis is rejected and the process is repeated for the hypothesis with the next highest confidence value in $H(I_q)$. If none of the hypotheses overcomes the belief threshold the image is categorized based on the maximum confidence degree of $H(I_q)$. One disadvantage of this approach lies on the difficulty of estimating an optimal belief threshold adapted to the statistical characteristics of each hypothesis. However, the fact that only a small portion of the available classifiers is required to reach a decision makes this approach attractive for complex domains.

IV. Ontology to Bayesian Network mapping

A BN is a directed acyclic graph $G = (V, A)$ whose nodes $v \in V$ represent variables and whose arcs $a \in A$ encode the conditional dependencies between them. Using the Bayes theorem, and given that a subset of variables are observed, the marginal probabilities of the remaining variables in the network can be estimated. The reason for using BNs in our framework is to estimate the posterior probabilities $\acute{H}(I_q)$ of the concepts in the hypothesis set c^H, using the observed confidence degrees $E(I_q)$ of the concepts in the evidence set c^E. However, given that the network structure is capable of encoding the qualitative characteristics of causality (i.e., which nodes affect which), and the Conditional Probability Tables (CPTs) can be used to quantify the causality relations between concepts (i.e., how much is a node influenced by the nodes to which it is connected), the constructed BN will be able to facilitate three different operations: a) Provide the means to store and utilize domain knowledge K_D; this is achieved by mapping K_D to the network structure. b) Organize and make accessible information coming from the application context W_{ij}; this is achieved by the CPTs attached to the network nodes. c) Allow the propagation of evidence belief in a mathematically coherent manner; this is performed with the use of message passing belief propagation algorithms. The work of [30] describes a probabilistic extension to OWL ontology based on BNs and define a set of structural translation rules to convert this ontology into a directed acyclic graph. Here, we propose an adaptation of this method that learns the network parameters from data in contrast to being explicitly defined by an expert.

A. Network Structure

Deciding on the structure of a BN based on an ontology can be seen as mapping ontological elements (i.e., concepts and relations) to graph elements (i.e., nodes and arcs). Thus, we have $S(N_C, R, O) \to G(V, A)$, with $N_C \to V$, $R \to A$, and $O \to (V, A)$. The symbol $O \to (V, A)$ indicates that in order to migrate a DL constructor into the network structure both nodes and arcs will have to be employed. The structural transformation process adopted in our framework takes place in two stages. In the first stage, the BN incorporates the hierarchical information of the ontology. In order to do so, all ontology concepts are transformed into network nodes with two states (i.e., true and false). These nodes are called concept nodes n_{cn}. Then, an arc is drawn between two concept nodes in the network, if and only if they are connected with a superclass-subclass relation in K_D and with the direction from the superclass to the subclass node. The adoption of this principle was motivated by the fact that when an instance belongs to a certain class it is automatically subsumed that it can also belong to one of its subclasses, thus imposing a kind of causality. At the second stage, the BN incorporates the semantic constraints between concepts that are expressed in the ontology. This is done by creating a control node n_{cl} for each DL constructor, which is connected to the concept nodes that correspond to the concepts associated with this constructor. The way in which the connection is made depends on the type of the DL constructor and results in a different sub-network structure, see [30] for details. The DL constructors that can be handled by the adopted methodology are owl:intersectionOf, owl:unionOf, owl:complementOf, owl:equivalentClass and owl:disjointWith.

B. Parameter Learning

Once the network structure is fixed, each concept node n_{cn} needs to be assigned a prior probability if it is a root node or a conditional probability table if it is a child of one or more nodes. In [30] these probabilities are set by domain experts and formulate the original probability distribution of the network. In order to learn the probability distribution of the network enhanced with the semantic constraints of the domain, the authors developed the D-IPFP algorithm, which is an algorithm based on the “iterative proportional fitting procedure” (IPFP). This procedure modifies a given distribution to meet a set of constraints (i.e., the semantic constraints of the domain), while minimizing the *KL-divergence* (Kullback-Leibler divergence) to a target distribution (i.e., the original probability distribution of the network). The drawback of this approach is that apart from requiring human intervention when switching to a different domain, it is also likely to introduce bias in the initial conditions of the BN.

In our work, we propose a variation of the aforementioned methodology where the original probability distribution is learned from sample data instead of being explicitly provided by humans. The sample data are concept labels that have been used to annotate the image dataset at both global and region level. Given a sufficiently large amount of annotated images, the original probability distribution of the network can be approximated using the frequency information implicit in the data. Such an approach is frequently employed by works that use graph-based probabilistic networks [40], [41], where in contrast to [30] the conditional probabilities are learned using a sample portion of the data that is being modeled. However, learning from data can only be done robustly if there is a sufficiently large amount of samples available. In any other case, as will become clear in Section VI-D, the estimated conditional probabilities are inaccurate and tend to mislead the inference process.

The conditional probabilities are learned by employing the Expectation Maximization (EM) [42] algorithm, using as training data the images annotated with concept labels. Initially, we apply the EM algorithm to a BN that incorporates only the hierarchical information of the ontology. Then, we add the control nodes to model the semantic constraints and we once again apply the EM algorithm to the modified BN. Since no sample data are available for the control nodes, these nodes are treated as latent variables with two states (i.e., true and false). The last step is to manually set the CPTs of all control nodes n_{cl} as shown in [30] and set the belief of the true state equal to 100%. This is done in order to enforce the semantic constraints into the probabilistic inference process.

V. Framework functional settings

A. Image Analysis Tasks

This section describes how the proposed framework can be adapted to three different image analysis tasks. For each of these tasks we clarify the task specific contextual information $app \in X$ (i.e., where to physically search for evidence) as well as the way that the hypotheses $H(I_q)$ and evidence $E(I_q)$ sets are determined.

Image categorization is the task of selecting the category concept c_i that best describes an image I_q as a whole. In this case, a hypothesis is formulated for each of the category concepts, that is $H(I_q) = \{Pr(c_i|I_q) : i = 1, \ldots, n\}$ where n is the number of category concepts in K_D. Global classifiers (i.e., models trained using global image information) are applied to estimate the initial probability for each hypothesis. For this task, the application context app determines which evidence should be taken from the image regions extracted using a segmentation algorithm. For instance, knowing that a specific region depicts *road* is a type of contextual information that the algorithm can exploit when trying to decide whether the image depicts a *Seaside* or a *Roadside* scene. Local classifiers (i.e., models trained using regional image information) are applied to the pre-segmented image regions $I_q^{s_j}$, in order to generate a set of confidence values that constitute the evidence $E(I_q) = \{Pr(\acute{c}_i|I_q^{s_j}) : i = 1, \ldots, k \quad \& \quad j = 1, \ldots, m\}$, where k is the number of regional concepts in K_D and m is the number of identified segments. In this case, the category concepts c_i constitute the hypothesis set c^H and the regional concepts $\acute{c}_i$ comprise the evidence set c^E.

Localized region labeling, is the task of assigning labels to pre-segmented image regions, with one of the available regional concepts $\acute{c}_i$. In this case, a hypothesis is formulated for each of the available regional concepts and for each of the image segments. That is $H(I_q) = \{Pr(\acute{c}_i|I_q^{s_j}) : i = 1, \ldots, k \quad \& \quad j = 1, \ldots, m\}$, where k is the number of regional concepts and m is the number of identified segments. Local classifiers are used to estimate the initial probability for each of the formulated hypotheses. In this task, the contextual information app is considered to be the image as a whole. For example, knowing that an image depicts a *Roadside* scene can be considered the application context and facilitate the algorithm to decide whether a specific region depicts *sea* or *road*. The degrees of confidence for each of the category concepts c_i, obtained by applying the global classifiers to I_q, constitute the evidence of this task. That is $E(I_q) = \{Pr(c_i|I_q) : i = 1, \ldots, n\}$, where n is the number of category concepts. In this case, the regional concepts $\acute{c}_i$ constitute the hypothesis set c^H and the category concepts c_i comprise the evidence set c^E.

In practice, our framework can be used to improve region labeling when there is a conflict between the decisions suggested by the global and local classifiers. A conflict occurs when the concept suggested by the local classifiers does not belong to the set of child nodes of the concept suggested by the global classifiers. Since there is no reason to trust one suggestion over another we make two different hypotheses. The first one assumes that the suggestion of the global classifiers is correct. The regional concept corresponding to the maximum confidence degree, among the child nodes of the category concept, is selected and the overall impact on the posterior probability of the regional concept is measured. The second approach considers that the suggestion of the local classifiers is correct. The category concept corresponding to the maximum confidence degree, among the parent nodes of the regional concept suggested by the local classifiers, is selected and the overall impact on the posterior probability of the regional concept is measured. Among the two cases, the regional concept with the maximum positive impact on its posterior probability is selected to label the examined region.

Weak annotation of video shot key-frames is the task of associating a number of concepts with an image. However, in this case, we do not associate concepts with specific image regions. Thus, there is no distinction between category and regional concepts and more than one labels can be assigned to the image. A hypotheses set is formulated, $H(I_q) = \{Pr(c_i|I_q) : \quad i = 1, \ldots, n\}$ where n is the number of all available concepts in the domain. All classifiers are employed to extract the initial probability for all formulated hypotheses. The application context app determines that evidence should be searched for in the global image information. For instance, if an image is being examined for the presence of the concept *sports*, it would be helpful for the algorithm to know that the concept *soccer-player* is also depicted in the image. Thus, the evidence are considered to be the confidence values of all other concepts except the one examined by the current hypothesis. That means that when we examine the hypothesis $H(c_k|I_q)$, the evidence are $E(I_q) = \{Pr(c_i|I_q) : \forall i \in [1, n] \backslash \{k\}\}$.

B. Low-level Image Processing

The low level processing of visual stimulus consists of visual features extraction, segmentation and learning the concept detection models. Four different visual descriptors proposed in the MPEG-7 standard [43], namely Scalable Color, Homogeneous Texture, Region Shape, and Edge Histogram, were employed as described in [31]. Segmentation was performed using an extension of the Recursive Shortest Spanning Tree algorithm [44] that produces a segmentation mask $S = \{s_i : i = 1, \ldots, m\}$ for each image, with s_i representing the identified segment. Support Vector Machines (SVMs) were employed for learning the concept detection models (represented by F_c in Table I). Global and local classifiers were created off-line using manually annotated images as training samples and for all concepts included in K_D. The feature space is determined by the utilized visual descriptors and a gaussian radial basis is used as the kernel function.

For the task of weakly annotating video shot key-frames, we have utilized the detectors released by Columbia University [45]. In this case, individual SVMs were trained at global level independently over each feature space and a simple late fusion mechanism was subsequently applied to produce the average score. Three types of features were used, namely grid color moments, edge histogram direction and texture [45]. In all cases the SVM-based models were constructed using the libsvm library [46] and their soft output (i.e., confidence degree) was calculated based on the distance between the decision boundary and the classified feature vector in the kernel space. More specifically, the sigmoid function $Pr(c|I_q) = \frac{1}{1+e^{-td}}$ [47] was employed to compute the respective degree of confidence for a concept c, with t being a scale factor.

VI. Experimental Study

We present results for two datasets with different domain complexity and volume, namely the "Personal Collection" (*PS*) and the "News" (*NW*). *PS* was assembled internally in our lab by merging various photo albums while *NW* was taken from the TRECVID 2005 competition. Our goal is to demonstrate the improvement in performance achieved by exploiting context and knowledge compared to baseline detectors that rely solely on low-level visual information. We also evaluate the proposed FoA mechanism and show that we can significantly reduce the computational cost of visual inference and still achieve performance comparable to the exhaustive case. All experiments were conducted using the Netica software for handling BNs and the Prot$\acute{e}$g$\acute{e}$ ontology editor for constructing the ontologies.

A collection of 648 images I^{PS} comprised the dataset for the *PS* domain. All images in I^{PS} are annotated at global and region detail using the set of category concepts C_G={*Countryside_buildings, Seaside, Rockyside, Forest, Tennis, Roadside*} and the set of regional concepts C_L={*Building, Roof, Tree, Stone, Grass, Ground, Dried-plant, Trunk, Vegetation, Rock, Sky, Person, Boat, Sand, Sea, Wave, Road, Road-line, Car, Court, Court-line, Board, Gradin, Racket*}, respectively. For the *NW* domain 374 semantic concepts were defined by the Columbia University [45] to characterize its content. For this domain the TRECVID2005 development data [48] containing 137 annotated video clips were used. The annotations were provided at the level of subshots, extracted using temporal criteria (see [45] for details). By extracting a key-frame from each subshot, a dataset consisting of 61600 still images I^N annotated at global level was constructed.

In both cases, an ontology was used to represent domain knowledge. The ontology and the corresponding BN for the *PS* domain are depicted in Figs. 2 and 3, respectively. For the *NW* domain the ontology was constructed using the guidelines of [49]. More specifically, the concepts were associated on the basis of program categories N_G={*politics, finance/bussiness, science/technology, entertainment, weather, commercial/advertisement*} that were placed at the top of the hierarchy, having the rest of the concepts N_L as subclasses. Subsequently, the methodology of Section IV was applied to construct the corresponding BN. Both the ontology and the BN of the *NW* domain can be accessed through our web page [50].

I^{PS} was split in half to formulate the test I^{PS}_{test} and training I^{PS}_{train} sets, each one containing 324 images. I^{PS}_{train} was used for training the classifiers F_c and learning the parameters

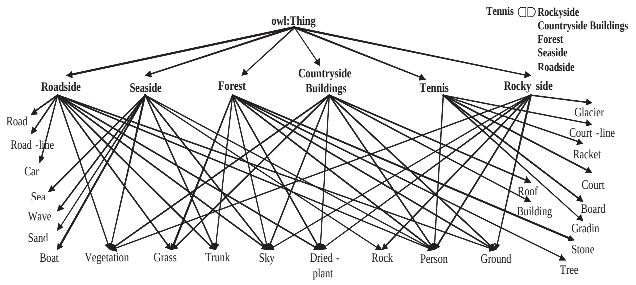


Fig. 2. Ontology encoding the domain knowledge about the "Personal Collection" domain.

of the BN. In a similar fashion, out of 137 video clips for the *NW* domain, the key-frames included in the first 100 I^{N}_{train} (i.e., 45276 still images) were selected for learning the parameters of the BN. The key-frames of the remaining 37 video clips I^{N}_{test} (i.e., 16624 still images) were used for testing. Concerning the classifiers, the baseline detectors of [45] were employed for all 372 concepts.

A. Image Categorization

We examine the efficiency of categorizing the images of I^{PS}_{test} to one of the categories in C_G using three configurations. These configurations vary in the amount of utilized context and knowledge. In the baseline configuration $CON1$ we assess the performance of image categorization based solely on visual stimulus. Images are categorized based on the maximum value of the global concept classifiers. The second configuration $CON2$ uses context (i.e., $X = S(app, W)$) and knowledge (i.e., $K_D = S(N_C, R, O)$) in order to extract the existing evidence and facilitate the process of evidence driven probabilistic inference. In this case, information from the image regions is incorporated into the analysis process but no semantic constraints are taken into account. The BN employed in this configuration is the one depicted in Fig. 3 without the nodes enclosed by the black frame. The joint probability distribution (JPD) of the random variables that are included in the BN utilized by $CON2$ configuration is:

$$Pr(C^1_G, .., C^{|G|}_G, C^1_L, .., C^{|L|}_L) = \prod_{i=1}^{|G|} Pr(C^i_G) \prod_{j=1}^{|L|} Pr(C^j_L | F(C^j_L)) \tag{3}$$

where $F(C^j_L)$ is the set of parent nodes of C^j_L according to the directed acyclic graph of the BN. The fact that none of the category concepts C_G has parent nodes (as shown in Fig. 3) allows us to include in the expression of the JPD the first product on the right hand side of eq. (3), which represents the product of the marginal probabilities of the category concepts. The third configuration $CON3$ takes into account the semantic constraints of the domain using the methodology presented in Section IV to construct the BN. In this case, the BN used for performing probabilistic inference is extended with the addition of the control nodes (i.e., the set of nodes enclosed by the black frame of Fig. 3) that are used for modeling the disjointness between *Tennis* and all other category concepts. If we define C_D to be the set of control nodes, the JPD defined by the BN utilized in $CON3$ configuration is:

$$Pr(C^1_G, .., C^{|G|}_G, C^1_L, .., C^{|L|}_L, C^1_D, .., C^{|D|}_D) = \prod_{i=1}^{|G|} Pr(C^i_G) \prod_{j=1}^{|L|} Pr(C^j_L | F(C^j_L)) \prod_{k=1}^{|D|} Pr(C^k_D | C^k_G, C^{Tennis}_G) \tag{4}$$

The use of the common superscript k in both C_D and C_G indicates that every node of the subnetwork that is used to model the disjointness between each category concept and *Tennis*, is conditioned on the node of the corresponding category concept and the node corresponding to *Tennis*. The reason for treating $CON2$ and $CON3$ as two different configurations was to examine how much of the overall improvement comes from the use of regional evidence and concept hierarchy information ($CON2$), and how much comes from the enforcement of semantic constraints in the analysis process ($CON3$).

In both $CON2$ and $CON3$ configurations the analysis process unfolds as follows. Initially, we formulate the hypotheses set using all category concepts. Then, we search for the presence of all possible regional concepts determined in K_D (i.e., $\forall c_j \in C_L$) before deciding which of them should be used as evidence. This approach requires the application of all available classifiers, global and local, for producing one set of confidence values for the image as a whole, $LK_{global} = \{Pr(c_i|I_q) : \forall c_i \in C_G\}$ (see Fig. 5, table with title "Global Classifiers") and one set per identified image region, $LK_{local} = \{Pr(c_j|I^{s_k}_q) : \forall c_j \in C_L \quad \& \quad \forall s_k \in S\}$. The latter is a matrix whose columns correspond to the regions identified by the segmentation algorithm and whose rows correspond to the confidence degrees of the regional concepts determined in K_D (see Fig. 5, table with title "Local Classifiers"). All values of LK_{global} and the maximum per column values of LK_{local} are introduced as soft evidence into the corresponding nodes of the BN. Then, the network is updated to propagate evidence impact and the concept corresponding to the node with the highest resulting posterior probability among the nodes representing category concepts, is selected to categorize the image (i.e., in this case $\otimes \equiv \max$, see Table I). Fig. 4 shows that the performance obtained using the $CON2$ is superior to the one obtained using $CON1$, since an average increase of approximately 5% is observed.

The running example of Fig. 5 demonstrates how evidence collected using regional information ($CON2$) can correct a decision erroneously taken by a global classifier that relies solely on visual stimulus ($CON1$). In Fig. 5, the Table "Global Classifiers" depicts the probabilities $Pr(c_i|I_q)$ that are obtained after the global classifiers are applied to image I_q. Using only this information, the image is categorized as *Seaside* (i.e., this is the result of $CON1$). *Seaside* is the chosen category even after inserting the values $Pr(c_i|I_q)$ into the network and performing inference (i.e., second row of table with title "*Belief Evolution*" in Fig. 5). However, as the pieces of regional evidence (i.e., the maximum value from each column of the "*Local Classifiers*" table), are inserted into the BN, belief propagation causes the posterior probabilities of the category concepts to change. The last four rows of "*Belief Evo-*

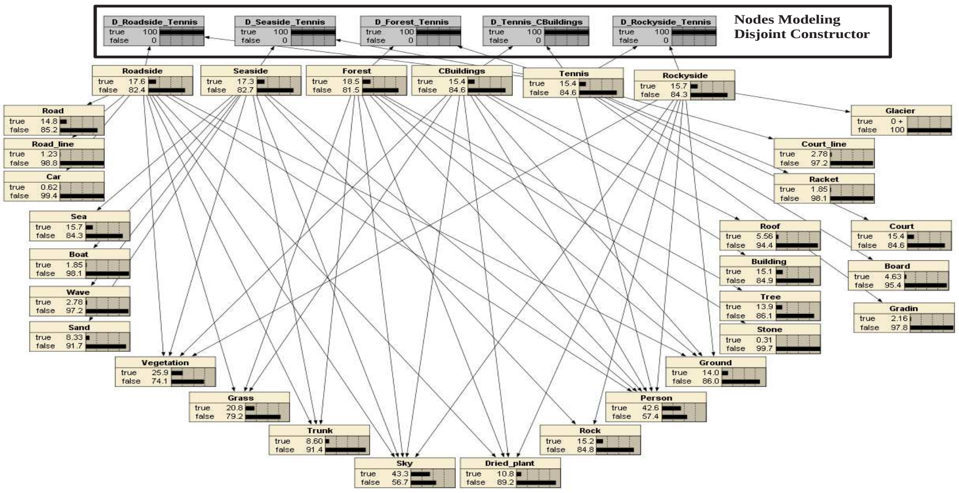

Nodes Modeling
Disjoint Constructor
D_Roadside_Tennis true 100 false 0
D_Seaside_Tennis true 100 false 0
D_Forest_Tennis true 100 false 0
D_Tennis_CBuildings true 100 false 0
D_Rockyside_Tennis true 100 false 0
Roadside true 17.6 false 82.4
Seaside true 17.3 false 82.7
Forest true 18.5 false 81.5
CBuildings true 15.4 false 84.6
Tennis true 15.4 false 84.6
Rockyside true 15.7 false 84.3
Road true 14.8 false 85.2
Road_line true 1.23 false 98.8
Car true 0.62 false 99.4
Glacier true 0 + false 100
Court_line true 2.78 false 97.2
Racket true 1.85 false 98.1
Sea true 15.7 false 84.3
Boat true 1.85 false 98.1
Wave true 2.78 false 97.2
Sand true 8.33 false 91.7
Roof true 5.56 false 94.4
Court true 15.4 false 84.6
Building true 15.1 false 84.9
Board true 4.63 false 95.4
Tree true 13.9 false 86.1
Gradin true 2.16 false 97.8
Stone true 0.31 false 99.7
Vegetation true 25.9 false 74.1
Ground true 14.0 false 86.0
Grass true 20.8 false 79.2
Person true 42.6 false 57.4
Trunk true 8.60 false 91.4
Rock true 15.2 false 84.8
Sky true 43.3 false 56.7
Dried_plant true 10.8 false 89.2

Fig. 3. Bayesian network derived from the ontology of Fig. 2 modeling the “Personal Collection” domain. The nodes in the black frame are control nodes that are used to model the disjointness between the concept *Tennis* and all other category concepts in the domain.

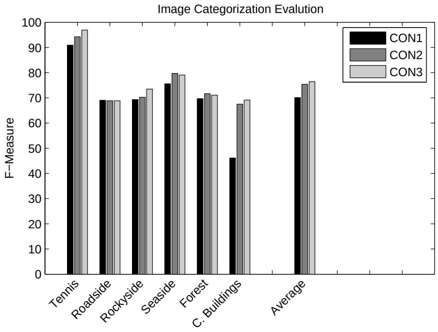
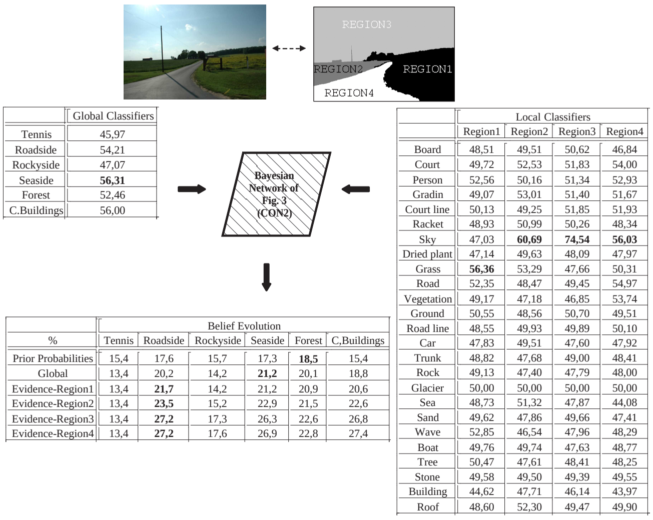


Fig. 4. F-Measure scores for the task of image categorization using CON1: the output of the global concept classifiers is used to categorize the image, CON2: uses also knowledge and application context for categorizing the image, and CON3: takes also into account the semantic constraints expressed in an ontology.

lution” table illustrate how the posterior probabilities evolve in the light of new evidence. Eventually the correct category, which is *Roadside*, emerges as the one with the highest posterior probability. It is interesting to note that only two out of four local classifiers (the ones corresponding to regions 1 and 3) predicted correctly the regional concept. Nevertheless, this information was sufficient for our framework to infer the correct prediction, since the relation between the concepts *grass* (identified in region 1) and *Roadside* was strong enough to raise the inferred posterior probability of this category above the corresponding value of *Seaside*. This is a reasonable result since the *Seaside* category receives no support from the evidence *grass*, as shown in Fig. 2.

The lower of cells in Table II depict the confusion matrix of $CON2$. By looking at the relations between regional and category concepts in Fig. 2 in conjunction with Table II, it is clear that our framework tends to confuse categories that share many regional evidence. This is the case for *Rockyside* and *Forest* or *Countryside Buildings* and *Roadside*. Another interesting observation is the small amount of regional evidence that *Tennis* shares with the rest of image categories. This can be practically considered as domain information (i.e., semantic constraint) and used to aid image analysis. In order to do so, we associate the *Tennis* concept and all other concepts in C_G with the “*owl:disjointWith*” DL-constructor. Then, we re-construct the BN using the enhanced ontology. The nodes of the BN that are enclosed by the black frame in Fig. 3 are used to model the disjointness between *Tennis* and all other category concepts. We can see from Fig. 4, that using the semantic constrains (CON3) the performance of image analysis is further increased with an average improvement of approximately 6.5%, compared to the baseline configuration ($CON1$). By inspecting the upper of the cells in Table II, where the confusion matrix for the $CON3$ is depicted, we can see that the improvement comes mainly from the correction of the test samples that were mis-categorized as *Tennis*.

In order to examine the statistical significance of this



	Global Classifiers
Tennis	45,97
Roadside	54,21
Rockyside	47,07
Seaside	**56,31**
Forest	52,46
C.Buildings	56,00

	Local Classifiers			
	Region1	Region2	Region3	Region4
Board	48,51	49,51	50,62	46,84
Court	49,72	52,53	51,83	54,00
Person	52,56	50,16	51,34	52,93
Gradin	49,07	53,01	51,40	51,67
Court line	50,13	49,25	51,85	51,93
Racket	48,93	50,99	50,26	48,34
Sky	47,03	**60,69**	**74,54**	**56,03**
Dried plant	47,14	49,63	48,09	47,97
Grass	**56,36**	53,29	47,66	50,31
Road	52,35	48,47	49,45	54,97
Vegetation	49,17	47,18	46,85	53,74
Ground	50,55	48,56	50,70	49,51
Road line	48,55	49,93	49,89	50,10
Car	47,83	49,51	47,60	47,92
Trunk	48,82	47,68	49,00	48,41
Rock	49,13	47,40	47,79	48,00
Glacier	50,00	50,00	50,00	50,00
Sea	48,73	51,32	47,87	44,08
Sand	49,62	47,86	49,66	47,41
Wave	52,85	46,54	47,96	48,29
Boat	49,76	49,74	47,63	48,77
Tree	50,47	47,61	48,41	48,25
Stone	49,58	49,50	49,39	49,55
Building	44,62	47,71	46,14	43,97
Roof	48,60	52,30	49,47	49,90

	Belief Evolution					
%	Tennis	Roadside	Rockyside	Seaside	Forest	C,Buildings
Prior Probabilities	15,4	17,6	15,7	17,3	**18,5**	15,4
Global	13,4	20,2	14,2	**21,2**	20,1	18,8
Evidence-Region1	13,4	**21,7**	14,2	21,2	20,9	20,6
Evidence-Region2	13,4	**23,5**	15,2	22,9	21,5	22,6
Evidence-Region3	13,4	**27,2**	17,3	26,3	22,6	26,8
Evidence-Region4	13,4	**27,2**	17,6	26,9	22,8	27,4

Fig. 5. Running example of image categorization using the framework’s $CON2$ configuration. The evidence extracted from image regions help to correct a misclassification error about the image category.

TABLE II
CONFUSION MATRIX FOR IMAGE CATEGORIZATION - $CON2$ LOWER OF THE CELLS - $CON3$ UPPER OF THE CELLS

%	Tennis	Roadside	Rockyside	Seaside	Forest	C. Buildings
Tennis	98.00 94.00	0.00 0.00	0.00 2.00	2.00 4.00	0.00 0.00	0.00 0.00
Roadside	1.75 0.00	73.68 73.68	0.00 0.00	8.77 8.77	10.53 12.28	5.26 5.26
Rockyside	5.88 0.00	3.92 3.92	64.71 70.58	5.88 5.88	19.61 19.61	0.00 0.00
Seaside	0.00 0.00	5.36 5.36	3.57 3.57	91.07 91.07	0.00 0.00	0.00 0.00
Forest	0.00 0.00	10.00 10.00	8.33 8.33	10.00 10.00	71.67 71.67	0.00 0.00
C. Buildings	2.00 0.00	24.00 24.00	6.00 6.00	12.00 12.00	2.00 2.00	54.00 56.00

improvement we apply the McNemar test on the output of $CON1$ and $CON3$ configurations. The 2×2 contingency table summarizing the transitions observed before and after employing our framework is depicted in Table III. Since the number of discordant pairs $(30 + 15)$ is more than 25, the chi-squared approximation with Yates' correction and 1 degree of freedom is calculated to be 4.536. Thus, the $p - value$ calculated by the McNemar's test equals 0.0369. By adopting the conventional criteria on statistical significance that considers the significance level α to be 0.05, we have $p - value < \alpha$. Thus, it is safe to conclude that the introduced improvement is statistically significant.

TABLE III
CONTINGENCY MATRIX - IMAGE CATEGORIZATION

		before		
		+	-	**Total**
after	+	218	30	**248**
	-	15	61	**76**
	Total	**233**	**91**	**324**

B. Image Categorization using a Focus of Attention Mechanism

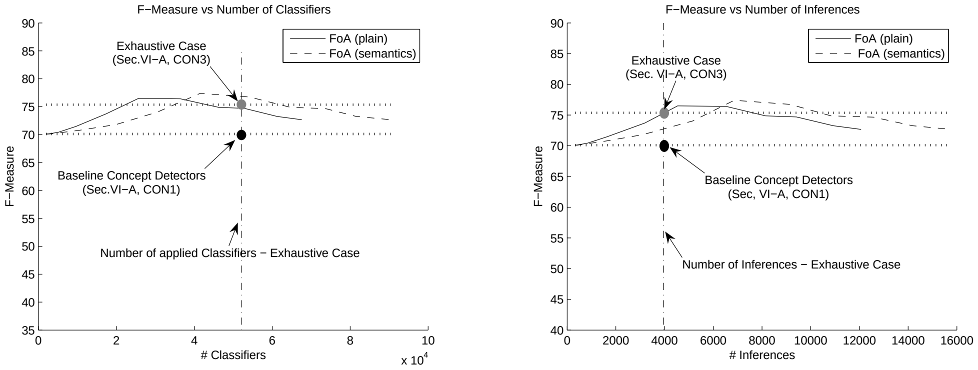
In order to assess the benefit of using the proposed Focus of Attention (FoA) mechanism, we measure the gain in computational cost in terms of two quantities. The number of classifiers (#Classifiers) that need to be applied and the number of inferences (#Inferences) that need to be performed. #Inferences is the number of times a confidence degree is inserted into one of the BN nodes and as a result triggers an inference process. When the FoA mechanism is not employed, the #Inferences that need to be performed for analyzing a single image is equal to the number of confidence values estimated for the global concepts of the image (i.e., the 6 values of LK_{global} in our experiments) plus the number of regions identified in the image (i.e., maximum per column values of LK_{local}). Thus, the total #Inferences for the complete set of 324 test images is $324 * 6$ plus the number of regions identified in all 324 test images, which was calculated to be 2010. Table IV shows the #Classifiers and #Inferences for the exhaustive case of Section VI-A (i.e., $CON3$). These values will serve as the baseline reference when estimating the computational gain of the FoA mechanism.

TABLE IV
COMPUTATIONAL COST QUANTITIES - $CON3$ CONFIGURATION

	324 (# Test Images) * 6 (# Global Classifiers) + 2010 (# Total Regions) * 25 (# Local Classifiers)
# Classifiers	**52194**
	324(# Test Images) * 6 (# Global Classifiers) + 2010 (max of local classifiers per region)
# Inferences	**3954**

In our experimental setting the belief threshold receives one of the following discrete values $\{0.1, 0.2, \ldots, 1.0\}$. Using each of these values as a common belief threshold for all formulated hypotheses, we obtain 10 different F-Measure scores. Given that the belief threshold affects also the #Classifiers and the #Inferences, we practically obtain 10 pairs of values for $\{$F-Measure, #Classifiers$\}$ and 10 pairs of values for $\{$F-Measure, #Inferences$\}$. These pairs are used to draw the curves depicted in Figs. 6(a) and 6(b). In both diagrams we demonstrate the performance of a) the baseline concept detectors (i.e., $CON1$ of Section VI-A) (black dot), b) the probabilistic inference using exhaustive search (i.e., $CON3$ of Section VI-A) (gray dot), c) the plain FoA mechanism (solid curve), and d) the FoA mechanism using also the methodology of Section IV for incorporating semantic constraints (dashed curve). The baseline figures of Table IV are also displayed in Figs. 6(a) and 6(b) using the vertical lines. The horizontal dotted lines are drawn for allowing comparisons with the performance of the baseline configurations. It is clear that the proposed FoA mechanism manages to achieve (for the optimal value of the belief threshold, F-Measure = $76, 40$) performance comparable to the one obtained by the best of the configurations in Section VI-A, using a remarkably smaller number of classifiers. On the other hand, for the same optimal threshold value, the number of inferences that need to be performed increases, see Fig. 6(b). More specifically, the number of classifiers reduces from 52194 to 25753 (# classifiers corresponding to the peak of the solid curve in Fig. 6(a)), while the number of inferences increases from 3954 to 4538 (# inferences corresponding to the peak of the solid curve in Fig. 6(b)). For the case where the FoA mechanism incorporates semantic constraints (dashed curve), the number of applied classifiers reduces from 52194 to 41560 (# classifiers corresponding to the peak of the dashed curve in Fig. 6(a)), while the number of inferences increases from 3954 to 6860 (# inferences corresponding to the peak of the dashed curve in Fig. 6(b)).

In order to estimate these numbers in terms of time we have calculated the average time per classifier and per inference to be $0, 12$ (sec) and $0, 69 * 10^{-3}$ (sec), respectively. Thus. the gain in computational time is approximately 3172 (sec) using the plain FoA mechanism and 1274 (sec) using the FoA with semantic constraints, which can be considered as a significant reduction of the overall computational cost. Finally, let us note that in both approaches for image categorization (Section VI-A and Section VI-B) the configuration incorporating semantic constraints outperforms the other configurations. This provides an additional argument for the effectiveness of the methodol-



(a) (b)

Fig. 6. F-Measure scores using the Focus of Attention mechanism against: a) # Classifiers, b) # Inferences. Each point in a curve corresponds to a belief threshold that receives one of the following discrete values {0.1, 0.2, ..., 1.0}.

ogy presented in Section IV.

C. Localized Region Labeling

In order to evaluate the performance of our framework for the task of assigning labels to pre-segmented regions, we have used the BN of Fig. 3 (without the nodes enclosed by the black frame) and the JPD of eq. (3). As mentioned in Section V-A, our framework can reinforce region labeling when there is a conflict between the decisions suggested by the global and local classifiers. Let $Child(c_k) = \{c_j : k \rightarrow_{parent} j\}$ be the subset of C_L corresponding to the child nodes of $c_k \in C_G$. Let also $LK_{global} = \{Pr(c_i|I_q) : \forall c_i \in C_G\}$ be the set of confidence values obtained from the global classifiers applied to image I_q and $LK^{s_w}_{local} = \{Pr(c_j|I^{s_w}_q) : \forall c_j \in C_L\}$ be the set of confidence values obtained from the local classifiers applied to a region $I^{s_w}_q$ of the image. A conflict occurs when $c_l \notin Child(c_g)$ with $g = \arg\max_i(LK_{global})$ and $l = \arg\max_j(LK^{s_w}_{local})$.

In the first case we follow the suggestion of the global classifiers and select the concept c_g. Then, the local concept c_l is selected such that $l = \arg\max_j(LK^{s_w}_{local})$ and $c_l \in Child(c_g)$. The confidence values corresponding to c_g and c_l are inserted into the BN as evidence and the overall impact on the posterior probability of the hypothesis stating that the region under examination $I^{s_w}_q$ depicts c_l is measured. In the second case, we follow the suggestion of the local classifiers and select $c_{\acute{l}}$, such that $\acute{l} = \arg\max_j(LK^{s_w}_{local})$. The confidence values of the global classifiers are examined and the $c_{\acute{g}}$ with $\acute{g} = \arg\max_i(LK_{global})$ and $c_{\acute{g}} \in F(c_{\acute{l}})$ is selected. As in the previous case, the confidence values corresponding to $c_{\acute{l}}$ and $c_{\acute{g}}$ are inserted into the network and the overall impact on the posterior probability of the hypothesis stating that the examined region $I^{s_w}_q$ depicts $c_{\acute{l}}$ is measured. Eventually, the values representing the impact on the posterior probabilities of the two different cases are compared and depending on the largest value, c_l or $c_{\acute{l}}$ is chosen to label the region in question (i.e., this is the functionality of $\otimes$ operator depicted in Table I, for this task). If no conflict occurs, the concept corresponding to the local classifier with maximum confidence is selected. Fig. 7(a) shows that when using the proposed framework an average increase of approximately 4.5% is accomplished.

In order to apply the McNemar's test for this case we calculate the 2×2 contingency matrix depicted in Table V. The $p-value$ estimated by the McNemar's test is found to be less than 0.0001 showing that the improvement is statistically very significant, since $p - value << \alpha$.

TABLE V
CONTINGENCY MATRIX - LOCALIZED REGION LABELING

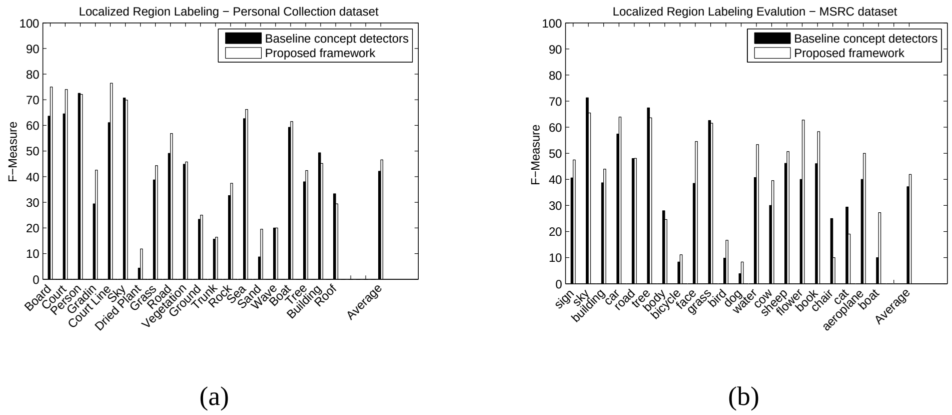
		before		
		+	-	**Total**
after	+	1035	61	**1096**
	-	22	892	**914**
	Total	**1057**	**953**	**2010**

D. Weakly Annotating Video Shot Key-Frames

This task does not require the existence of region-level annotations and therefore allows us to perform tests on a much larger set of semantic concepts. The TRECVID 2005 dataset was used for this purpose. Recalling that N_G denotes the set of category concepts that were placed at the top of the hierarchy and N_L the rest of domain concepts that were used as subclasses of N_G, the JPD defined by the utilized BN is:

$$Pr(N^1_G, .., N^{|G|}_G, N^1_L, .., N^{|L|}_L) = \prod_{i=1}^{|G|} Pr(N^i_G) \prod_{j=1}^{|L|} Pr(N^j_L|F(N^j_L)) \quad (5)$$

The benefit of using such a large dataset is the existence of semantic relations between the available concepts. These relations are necessary for assessing the effectiveness of our framework, since our goal is to exploit domain knowledge for improving the efficiency of image interpretation. On the other hand, many of the concepts appear rarely in the training set; a fact that makes difficult approximating the conditional probabilities using frequency information. In order to assess the efficiency of our framework we compare its performance against the performance of baseline concept detectors that make no use of domain knowledge and application context. In the first case we use the fused output of the global detectors released by the Columbia University [45]. The concepts



(a) (b)

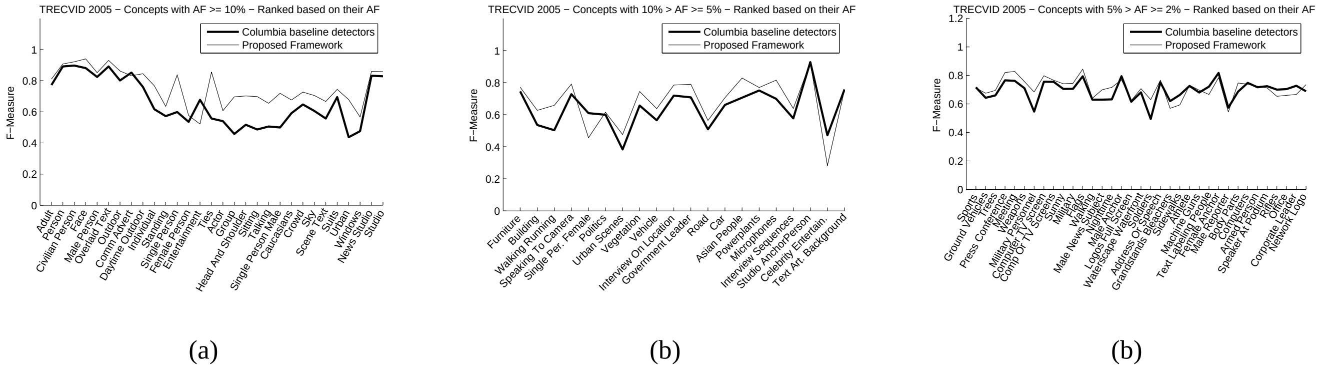
Fig. 7. F-Measure scores for the localized region labeling task: a) Personal Collection dataset, and b) Microsoft Research Cambridge dataset. Scores are reported for the baseline case, where decisions are based solely on the output of the classifiers, and for the proposed framework, where knowledge and context are employed to improve image analysis.

corresponding to the K maximum confidence values produced by the global detectors are selected to weakly annotate the key-frames. In the second case, the fused detection confidence values of all classifiers are provided as evidence to the BN. Belief propagation is performed and the resulting posteriors are recorded for all concepts. Finally, the K concepts that exhibit maximum positive impact on their posteriors were selected as the analysis outcome (i.e., this is the functionality of $\otimes$ operator depicted in Table I, for this task). For both cases, K was determined by varying its value between 2 and 20 and selecting the one that yields the optimal average F-Measure score.

In order to examine the relation between a concept's appearance frequency (AF) in the training set and the efficiency of the proposed framework, we report the F-Measure scores sorted based on the AF of the concepts. By inspecting Fig. 8(a) we observe that for the concepts with $AF \geq 10\%$ our framework outperforms the baseline in almost all cases. In Fig. 8(b), where the concepts with $10\% > AF \geq 5\%$ are depicted, we observe a similar behavior, but with the average improvement to be inferior from that of Fig. 8(a). Finally, Fig. 8(c) verifies that when the AF of a concept is relatively small (Fig. 8(c) depicts concepts with $5\% > AF \geq 2\%$) our framework does not deliver any improvements. Similar conclusions can be drawn when $AF < 2\%$. It is evident that the availability of realistic prior and conditional probabilities is important for the efficiency of our framework and learning them from data is feasible only when there are enough training samples to learn from.

E. Comparison with Existing Methods

In order to compare our work with other methods in the literature, we apply the localized region labeling task on the 591 images of the MSRC dataset [41] I^{MSRC}. In order to do so, we categorized all 591 images into 6 categories (i.e., global concepts), namely *Cityscape*, *Countryside*, *Forest*, *Indoors*, *ManMade* and *Waterside*. As regional concepts we used 21 out of the 23 semantic classes provided by MSRC, treating as void the *horse* and *mountain* classes that appear very rarely. An ontology was created to represent the relations between the aforementioned global and regional concepts and a BN was derived from it using the methodology presented in Section IV. Both of them can be accessed through our web page [50]. All images of I^{MSRC} were segmented and the ground truth label of each segment was considered to be the label of the hand-segmented region that overlapped with the segment by more than the 2/3 of the segment's area. In any other case the segment was labeled as void. We should note that although we could use directly the hand-segmented images of MSRC, such an approach would not be realistic since we cannot reasonably expect segmentation information for an unknown image. The overlap rule has been used by many works in the literature that utilize automatic image segmentation and need a way to decide the labels of the automatically extracted segments. For instance, [40] uses 20×20 image patches whose labels are considered to be the most frequent ground truth pixel label within the block, while [51] uses a 50% overlap rule between the segment's area and the ground truth foreground. The I^{MSRC} was split randomly in 295 training I^{MSRC}_{train} and 296 test I^{MSRC}_{test} images, ensuring approximately proportional presence of each class in both sets. I^{MSRC}_{train} was used from training the concept classifiers, as well as learning the parameters of the BN. Fig. 7(b) reports the performance for the baseline concept classifiers and the proposed framework configured as described in Section VI-C. The performance is increased in 14 out of the 21 regional concepts giving an average improvement of approximately 4.5%. The reason that some concepts like *sky*, *chair*, and *cat* exhibit performance lower from the baseline is the following. Our framework operates on top of the classifiers' outcome that usually come with a high number of erroneous predictions. Intuitively, the framework compensates for the misleading predictions by favoring the co-occurrence of evidence that are known from experience to usually co-exist and constitute the analysis context. It does so by adjusting the final output so as to comply with the extracted collection of evidence. Therefore, provided that an adequate amount of evidence are accurate, the framework is expected to make the correct decision by



(a) (b) (b)

Fig. 8. F-Measure scores for the concepts of TRECVID 2005 dataset ranked based on their appearance frequency (AF) in the training set: a) $AF \geq 10\%$, b) $10\% > AF > 5\%$, and c) $5\% > AF > 2\%$.

absorbing any misleading cues produced by the erroneous visual analysis. However, there can also be cases, like the ones mentioned above, where the evidence extracted from context are misleading, causing our framework to change the correct prediction of the local classifier.

In order to present results on the same dataset with [41] and [40], we have calculated the classification rate (i.e., number of correctly classified cases divided by the total number of correct cases) achieved by our framework for each of the 21 object classes in MSRC. We hereby note that the results depicted in Table VI are not directly comparable since they are reported at different levels. In [41] at pixel level, in [40] at the level of 20x20 image patches, and in our case at the level of arbitrary shaped, automatically extracted segments. In addition, the methods are not relying on the same set of visual features, and the training/test split is likely to be different.

It is clear that none of the approaches manages to outperform the others for a significant portion of the 21 classes. Moreover, error rates are often quite different on individual classes showing that while there are some classes that can be modeled very efficiently using the visual features and the model proposed by one method, there are other classes that are best modeled using a different set of visual features and model. For instance, while the visual features employed by our method perform very poorly in recognizing *grass*, they are pretty efficient in recognizing *car* or *sky*. Our aim is to use context and knowledge in order to improve the performance of a set of baseline concept classifiers by using their output as evidence, and not to discover a feature space that can best model an arbitrary set of classes.

VII. Discussion of the Results and Future Work

The conducted experiments verified the effectiveness of our framework in improving the performance of a set of baseline concept classifiers by using their output as evidence. Since this improvement derives mainly from the incorporation of the domain knowledge and the application context to the analysis process, we can use the proposed framework to improve the performance of any set of concept detectors that produce a probabilistic output. However, the results of the experiment in Section VI-A lead us to the conclusion that the amount and nature of the semantic information that can be used to enhance image interpretation depends largely on the special characteristics of the domain. More specifically, although using the information from the knowledge structure K_D and the causality relations $W_{ij} \in X$ obtained from context was proven to be useful in all cases, the semantic constraints originating from the domain were only able to facilitate image interpretation when the imposed rules were sufficiently concrete. For instance, the disjointness between "Tennis" and all other category concepts of the *PS* domain expresses a rather strict distinction that is suggested by knowledge. On the contrary, attempts to incorporate semantic constraints that, although valid from the point of logic, were less strict from the visual inference point of view didn't lead to performance improvements.

Furthermore, as shown in Section VI-D, a sufficiently large amount of training data is required for approximating the prior and conditional probabilities using frequency information. But given that the manual annotation of images is a cumbersome procedure, especially at region level, a solution could be to mine the necessary annotations from social sites like Flickr that are being populated with hundreds of user tagged images on a daily basis. Given that literature has already reported efforts on using this type of data [52], employing such schemes may help overcoming some of the problems caused from the use of limited size training sets.

Acknowledgment

This work was funded by the X-Media project (www.x-media-project.org) under EC grant number IST-FP6-026978 and the EC's Seventh Framework Programme FP7/2007-2013 under grant agreement 215453 - WeKnowIt.

References

[1] S.-F. Chang, "The holy grail of content-based media analysis," *IEEE MultiMedia*, vol. 9, no. 2, pp. 6–10, 2002.

[2] A. F. Smeaton, "Content vs. context for multimedia semantics: The case of sensecam image structuring," in *SAMT*, 2006, pp. 1–10.

[3] A. Oliva and A. Torralba, "Building the gist of a scene: the role of global image features in recognition," in *Progress in Brain Research*, 2006, p. 2006.

[4] D. Gokalp and S. Aksoy, "Scene classification using bag-of-regions representations," in *CVPR '07*, june 2007, pp. 1 –8.

[5] A. Opelt, A. Pinz, and A. Zisserman, "Incremental learning of object detectors using a visual shape alphabet," in *CVPR '06*, 2006, pp. 3–10.

TABLE VI
COMPARISON WITH EXISTING METHODS IN OBJECT RECOGNITION

	Buildings	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat	Average
Textonboost [41]	**62**	**98**	86	58	50	83	60	53	**74**	63	**75**	63	35	**19**	**92**	15	86	**54**	19	**62**	7	58
PLSA-MRF/P [40]	52	87	68	**73**	**84**	94	**88**	**73**	70	68	74	**89**	33	19	78	**34**	**89**	46	**49**	54	**31**	**64**
Prop. Fram.	32	55	**87**	40	73	**96**	57	56	50	**76**	8	64	**38**	12	46	5	51	12	8	29	18	44

[6] T. Blaschke, "Object-based contextual image classification built on image segmentation," in *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, 2003*, oct. 2003, pp. 113 – 119.

[7] Z. Wang, D. Feng, and Z. Chi, "Region-based binary tree representation for image classification," in *International Conference on Neural Networks and Signal Processing, 2003.*, vol. 1, 2003, pp. 232 – 235.

[8] J. Fan, Y. Gao, and H. Luo, "Multi-level annotation of natural scenes using dominant image components and semantic concepts," in *12th ACM international conference on Multimedia*, New York, USA, 2004, pp. 540–547.

[9] C. Yang, M. Dong, and F. Fotouhi, "Region based image annotation through multiple-instance learning," in *13th ACM international conference on Multimedia*, New York, NY, USA, 2005, pp. 435–438.

[10] T. Malisiewicz and A. A. Efros, "Recognition by association via learning per-exemplar distances," *CVPR '08*, pp. 1–8, 2008.

[11] P. Murphy, A. Torralba, and W. T. Freeman, "Using the forest to see the trees: a graphical model relating features, objects and scenes," in *in Advances in Neural Information Processing Systems (NIPS)*, 2003.

[12] A. Singhal, J. Luo, and W. Zhu, "Probabilistic spatial context models for scene content understanding," *CVPR '03*, p. 235, 2003.

[13] M. Boutell, J. Luo, and C. Brown, "Improved semantic region labeling based on scene context," in *ICME '05.*, 2005.

[14] L. Yang, P. Meer, and D. J. Foran, "Multiple class segmentation using a unified framework over mean-shift patches," *CVPR '07*, pp. 1–8, 2007.

[15] M. Boutell, J. Luo, and C. M. Brown, "A generalized temporal context model for classifying image collections," *Multimedia Syst.*, vol. 11, no. 1, pp. 82–92, 2005.

[16] L. Paletta, M. Prantl, and A. Pinz, "Learning temporal context in active object recognition using bayesian analysis," *International Conference on Pattern Recognition,*, vol. 1, p. 1695, 2000.

[17] D. Moldovan, C. Clark, and S. Harabagiu, "Temporal context representation and reasoning," in *IJCAI'05*, 2005, pp. 1099–1104.

[18] M. Boutell and J. Luo, "Bayesian fusion of camera metadata cues in semantic scene classification," in *CVPR '04*, vol. 2, 2004, pp. 623 –630.

[19] X. He, R. S. Zemel, and M. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," *CVPR '04*, vol. 2, pp. 695–702, 2004.

[20] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *ICCV '05*, 2005, pp. 1284–1291.

[21] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *ICCV '09*, Kyoto, Japan, September 2009.

[22] M. R. Naphade, T. T. Kristjansson, B. J. Frey, and T. S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems," in *ICIP '98*, 1998, pp. 536–540.

[23] M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 141–151, 2001.

[24] J. Luo, A. E. Savakis, and A. Singhal, "A bayesian network-based framework for semantic image understanding," *Pattern Recognition*, vol. 38, no. 6, pp. 919–934, 2005.

[25] P. Mylonas, E. Spyrou, Y. Avrithis, and S. Kollias, "Using visual context and region semantics for high-level concept detection," *IEEE Transanctions on Multimedia*, vol. 11, no. 2, pp. 229–243, 2009.

[26] M. J. Kane and A. E. Savakis, "Bayesian network structure learning and inference in indoor vs. outdoor image classification," in *ICPR '04*, 2004, pp. 479–482.

[27] L. N. Matos and J. M. de Carvalho, "Combining global and local classifiers with bayesian network," in *ICPR '06*, 2006, p. 952.

[28] S. Todorovic and M. C. Nechyba, "Interpretation of complex scenes using dynamic tree-structure bayesian networks," *Comput. Vis. Image Underst.*, vol. 106, no. 1, pp. 71–84, 2007.

[29] W. Liao and Q. Ji, "Learning bayesian network parameters under incomplete data with domain knowledge," *Pattern Recogn.*, vol. 42, no. 11, pp. 3046–3056, 2009.

[30] Z. Ding, Y. Peng, and R. Pan, "A bayesian approach to uncertainty modeling in owl ontology," in *International Conference on Advances in Intelligent Systems - Theory and Applications*, Nov. 2004.

[31] G. T. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Combining global and local information for knowledge-assisted image analysis and classification," *EURASIP J. Adv. Sig. Proc.*, vol. 2007, no. 2, pp. 18–18, 2007.

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

[33] J. Fan, Y. Gao, and H. Luo, "Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation," *IEEE Transactions on Image Processing*, vol. 17, no. 3, pp. 407–426, 2008.

[34] C. Town, "Ontological inference for image and video analysis," *Mach. Vis. Appl.*, vol. 17, no. 2, pp. 94–115, 2006.

[35] D. L. McGuinness and F. van Harmelen, "OWL web ontology language overview," W3C, W3C Recommendation, Feb. 2004.

[36] I. Horrocks, "Description logics in ontology applications," in *Automated Reasoning with Analytic Tableaux and Related Methods*, 2005, pp. 2–13.

[37] J. Cardoso, "The semantic web vision: Where are we?" *IEEE Intelligent Systems*, vol. 22, no. 5, pp. 84–88, 2007.

[38] J. A. Toth, *Reasoning agents in a dynamic world: The frame problem.*, ser. Artificial Intelligence, K. M. Ford and P. J. Hayes, Eds. Elsevier, 1995, vol. 73.

[39] F. V. Jensen and F. Jensen, "Optimal junction trees," in *Proc. of the 10th Conf. on Uncertainty in Artif. Intel.*, C. M. Kaufmann, Ed., San Mateo, 1994.

[40] J. Verbeek and B. Triggs, "Region classification with markov field aspect models," *CVPR '07*, pp. 1–8, 2007.

[41] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "*TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *ECCV '06*, 2006, pp. 1–15.

[42] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, 2nd ed. John Wiley and Sons, 1997.

[43] B. S. Manjunath, J. R. Ohm, V. V. Vinod, and A. Yamada, "Colour and texture descriptors," *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7*, vol. 11, pp. 703–715, 2001.

[44] N. M. T. Adamek, N. O'Connor, "Region-based segmentation of images using syntactic visual features," in *WIAMIS '05*, Montreux, Switzerland, 2005.

[45] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu, "Columbia universitys baseline detectors for 374 lscom semantic visual concepts," Columbia University, Technical Report 222-2006-8, 2007.

[46] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," 2001.

[47] D. Tax and R. Duin, "Using two-class classifiers for multiclass classification," in *ICPR '02*, Quebec, Canada, 2002.

[48] NIST, "Trec video retrieval evaluation (trecvid)," 2001-2006.

[49] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann, "A light scale concept ontology for multimedia understanding for trecvid 2005," IBM, Tech. Rep., 2005.

[50] S. Nikolopoulos. (2011) Public web page for auxiliary content. [Online]. Available: http://mklab.iti.gr/content/evidence-driven-image-interpretation-using-ontologies-bayesian-networks

[51] F. Ge, S. Wang, and T. Liu, "Image-segmentation evaluation from the perspective of salient object extraction," in *CVPR '06*, 2006, pp. 1146–1153.

[52] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their localization in images," in *ICCV*, 2005, pp. 370–377.

Spiros Nikolopoulos received his diploma degree in Computer Engineering and Informatics and the MSc degree in Computer Science & Technology from university of Patras, Greece in 2002 and 2004 respectively. He is currently working as a research associate with the Informatics and Telematics Institute (ITI) and he is a PhD candidate at Queen Mary University of London. His research interests include knowledge-based image analysis, content-based indexing and retrieval, cross-media information extraction and social media analysis. He has published 7 articles in international journals and books and 19 papers in international conferences and workshops. He currently serves as a reviewer of the IEEE community.

Georgios Th. Papadopoulos was born in Thessaloniki, Greece in 1982. He received the Diploma degree and the Ph.D. degree in Electrical and Computer Engineering from the Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece, in 2005 and 2011, respectively. His research interests include knowledge-assisted multimedia analysis, content-based and semantic multimedia indexing and retrieval, context-based semantic multimedia analysis, multimodal analysis, computer vision, pattern recognition and machine learning techniques. He has published 5 international journals and he is the coauthor of 19 papers in international conferences. He is a member of the IEEE and the Technical Chamber of Greece.

Dr. Ioannis (Yiannis) Kompatsiaris is a Senior Researcher (Researcher B) with the Informatics and Telematics Institute. His research interests include semantic multimedia analysis, indexing and retrieval, social media analysis, knowledge structures, reasoning and personalization for multimedia applications. He received his Ph.D. degree in 3-D model based image sequence coding from the Aristotle University of Thessaloniki in 2001. He is the co-author of 42 papers in refereed journals, 20 book chapters, 4 patents and more than 150 papers in international conferences. He has been the co-organizer of various international conferences and workshops and has served as a regular reviewer for a number of journals and conferences. He is a member of IEEE and ACM..

Ioannis (Yiannis) Patras (S' 1997, M'2002, SM' 2011) received the B.Sc. and M.Sc. degrees in computer science from the Computer Science Department, University of Crete, Heraklion, Greece, in 1994 and in 1997, respectively, and the Ph.D. degree from the Department of Electrical Engineering, Delft University of Technology, The Netherlands, in 2001. Between 2005 and 2007 was a Lecturer in Computer Vision at the Department of Computer Science, University of York, York, UK. He is a Senior Lecturer in Computer Vision in the School of Electronic Engineering and Computer Science in the Queen Mary, University of London. He is/has been in the organizing committee of IEEE SMC 2004, Face and Gesture Recognition 2008, ICMR2011, ACM Multimedia 2013 and was the general chair of WIAMIS 2009. He is associate editor in the Image and Vision Computing Journal. His research interests lie in the areas of Computer Vision and Pattern Recognition, with emphasis on Human Sensing and its applications in Multimedia Retrieval, Multimodal Human Computer Interaction. Currently, he is interested in the analysis of Human Motion, including the detection, tracking and understanding of facial and body gestures.