An evidence-driven probabilistic inference framework for semantic image understanding

Spiros Nikolopoulos[1], Georgios. Th. Papadopoulos[1], Ioannis Kompatsiaris[1], and Ioannis Patras[2]

[1] Informatics and Telematics Institute, CERTH
6th km Charilaou-Thermi Road, Thessaloniki, Greece
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, London, UK

Abstract. This work presents an image analysis framework driven by emerging evidence and constrained by the semantics expressed in an ontology. Human perception, apart from visual stimulus and pattern recognition, relies also on general knowledge and application context for understanding visual content in conceptual terms. Our work is an attempt to imitate this behavior by devising an evidence driven probabilistic inference framework using ontologies and bayesian networks. Experiments conducted for two different image analysis tasks showed improvement in performance, compared to the case where computer vision techniques act isolated from any type of knowledge or context.

1 Introduction

The use of knowledge and context for indexing multimedia data using higher level semantics, was motivated by the gap existing between the limited inference capabilities that restrain machine understanding and the plentiful reasoning potentials of human brain. Driven by the fact that knowledge and context are two dimensions of human perception that are difficult to introduce and exploit at the numeric level of visual features, we investigate the combined use of formal represented semantics and probabilistic inference mechanisms as a means to simulate their impact on image analysis. Evidence is information that when coupled with the principles of inference becomes relevant to the support or disproof of a hypothesis. For our framework visual stimulus is considered evidence when reasoned on the grounds of knowledge and placed on the appropriate context. In this perspective, the input arguments of an evidence-driven probabilistic inference framework consists of visual stimulus, application context and domain knowledge, as can be seen in Fig. 1. Application context and domain knowledge also affect the process of probabilistic inference (Fig. 1) and are considered to be the a priori/fixed information of the framework. On the other hand, the visual stimulus depends on the image to be analyzed and is considered to be the observed/dynamic information of the framework.

Domain knowledge, expressed using ontologies, and application context, captured both in conditional probabilities and application specific structures, are

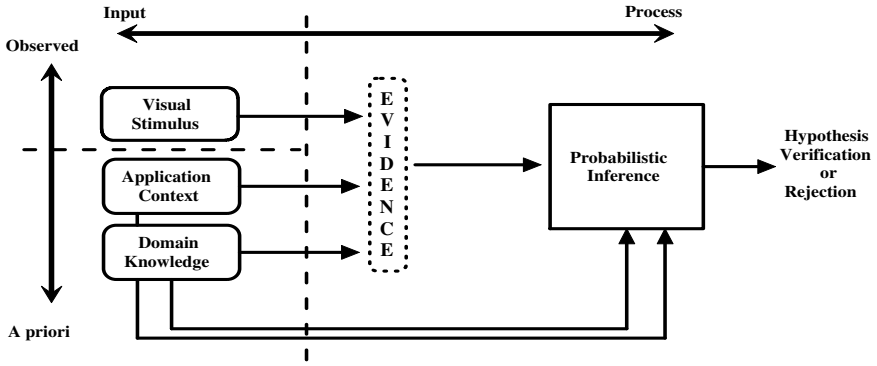


Fig. 1. Functional relations between the modules of the proposed framework.

integrated into a decision model that bears the characteristics of a bayesian network. A methodology allowing the automatic integration of ontology-expressed knowledge into a probabilistic network, is employed for this purpose. The goal of this network is to facilitate evidence driven probabilistic inference in order to verify or reject a hypothesis made about the semantic content of an image. Hence, the tasks carried out by the proposed framework include the statistical analysis of the presented visual stimulus, the adoption of a probabilistic standpoint for coherently handling uncertainty (a feature inherent to multimedia analysis), the representation of domain knowledge and application context (in the form of causality between evidence and hypotheses) in a computationally enabled format and the establishment of a framework supporting decision making driven by the probabilistic inferences of evidence. Thus, what can be considered as the contribution of our work is the fact that the potentials of such techniques i.e., techniques that integrate explicitly provided knowledge and bayesian networks, are thoroughly examined and evaluated as a means to enhance semantic image understanding by allowing in a principled/probabilistic way, the fusion of evidence/information obtained using knowledge and context.

2 Related Work

Various works exist in the literature that try to handle uncertainty and take advantage of knowledge and context for enhancing image analysis. Naphade et al. in [1] introduce the concept of "Multijects" as a way to map time sequence of multi-modal, low-level features to higher level semantics using probabilistic rules. In [2] Naphade et al. illustrates the functionality of "Multinets" by introducing bayesian belief networks as a means to model the interaction between concepts and use this contextual information for performing semantic indexing of video content. Luo et al. [3] propose a framework for semantic image understanding based on belief networks. The authors demonstrate the improvement in performance introduced by extracting and integrating in the same bayesian

inference framework, both low-level and semantic features. Other approaches that take advantage of knowledge and context include [4], [5], where indoor versus outdoor scene categorization based on low-level features and bayesian networks is performed and [6] where a bayesian network is utilized as an inference mechanism for facilitating a classification method. However, none of these works incorporate a systematic methodology for integrating domain knowledge, expressed with standard knowledge representation languages, into a probabilistic inference framework. On the other hand, Ding et al. in [7] present their on-going research on combining ontologies and bayesian networks with the aim of introducing uncertainty in ontology reasoning and mapping. However, no attempt is made by the authors to adjust their scheme for serving the purposes of multimedia analysis. In [8], Papadopoulos et al. propose 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 fuses intermediate classification results with contextual information and spatial relations, is used to generate the final results. In [9] Athansiadis et al. propose a scheme that is intended to enhance typical 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. However, no coupling of ontology-based approaches with probabilistic inference algorithms is attempted. Town in [10] use ontologies as a structural prior for deciding the structure of a bayesian network using the K2 algorithm. The task facilitated concerns the high-level analysis of surveillance data, but in this work ontologies are mostly treated as hierarchies that do not incorporate any explicitly represented semantic constraints.

3 Framework Description

What people see is not simply a translation of visual stimulus since knowledge and context have a major effect. The framework of Fig. 1 attempts to simulate visual perception by using evidence as an intermediate layer of image interpretation that combines visual stimulus, domain knowledge and application context.

Visual Stimulus: Machine learning methods are able to train a general classifier in recognizing a specific object, provided that a sufficiently large number of indicative examples are available. Thus, a classifier F can be trained to recognize a concept c based on the attributes f_I of its visual representation I. The output of such a classifier can be either binary or a value ranging between 0 and 1 that indicates the confidence (or uncertainty) of the classification output, i.e., $F_c(f_{I_q}) = Pr(c \mid I_q)$. $Pr(c \mid I_q)$ expresses the probability that visual representation I_q depicts concept c.

Domain Knowledge: Domain knowledge will have to be elucidated and represented in machine understandable format in order to be exploitable by our framework. Ontologies have emerged as a very powerful tool able to express knowledge in different levels of granularity [11]. If we consider N_C to be the set of

unary predicate symbols that are used to denote concepts, let R to be the set of binary predicates that are used to denote relations between concepts and O the algebra defining the allowable operands for these sets, the part of experience that relates to the domain knowledge can be represented using N_C, R, O. Thus, using OWL–DL [12], the domain knowledge can be expressed by a structure K_D that associates the domain concepts and relations using the allowable operands, i.e., $K_D = S(N_C, R, O)$, where $O \in DL$. DL stands for "Description Logics" [13] and constitutes a specific set of constructors and restrictions. For instance, common constructors include logical operands of the type *intersection*, *union*, *disjoint*, etc. Our goal is to use these constructors for imposing semantic constraints on the process of image interpretation that can not be captured by typical machine learning techniques.

Application Context: Loosely speaking the knowledge structure determines a) what evidence to look for, a piece of information that is associated with the domain knowledge and b) how to qualitatively evaluate their presence (i.e., which evidence supports one hypothesis or another). In this sense, the knowledge structure sets the tracks to which evidence belief is allowed to propagate. However, no support is provided to the decision making process in terms of where to look for evidence and how to quantitatively evaluate their presence (i.e., how much each hypothesis is affected by the existence of one evidence or another). The role of K_D is to capture information about the domain of discourse in general, not to deliver information concerning the context of the analysis process at hand. This is the role of application context that typically incorporate many application specific information. If we let *app* denote the set of application specific information (where to look for evidence in our case) and W_{ij} a function that quantifies the influence (i.e., measured as the frequency of co-occurrence) of concept c_i on c_j, the application context can be expressed as a structure of the type $X = S(app, W)$.

Evidence-driven Probabilistic Inference: An evidence-driven probabilistic inference scheme should be able to acquire what evidence to look for, from $N_C \in K_D$, use context information $app \in X$ to search for these evidence and apply the trained classifiers F_c to obtain the respective degrees of confidence. Subsequently, formulate an initial hypothesis around a concept $c \in N_C$ for all potential decisions, use the evidence to trigger probabilistic inference, propagate evidence beliefs using the inference tracks $R \in K_D$ and the corresponding belief quantification functions $W_{ij} \in X$, re-estimate the values for all hypotheses and finally decide which of the hypotheses should be verified or rejected. K_D and $app \in X$ determine which of the available concepts should be included in the hypotheses set c^H and which in the evidence set c^E. In this case, if we denote $H(I_q) = \{Pr(c_1^H \mid I_q), \ldots, Pr(c_M^H \mid I_q)\}$ the estimated degrees of confidence (i.e., prior probabilities) of the concepts belonging to the hypotheses set and $E(I_q) = \{Pr(c_1^E \mid I_q), \ldots Pr(c_K^E \mid I_q)\}$ the estimated degrees of confidence of the concepts belonging to the evidence set, evidence driven probabilistic inference is the process of calculating the posterior probabilities of $H(I_q)$ (i.e., $\acute{H}(I_q)$) given the evidence values $E(I_q)$ and information coming

from knowledge R, O and context W_{ij}. Thus, the proposed framework achieves semantic image interpretation in the following way, $c = \arg \otimes_{c^H} (\acute{H}(I_q))$ where $\acute{H}(I_q) = Pr(c^H \mid H(I_q), R, O, W_{ij}, E(I_q))$ and $\otimes$ is an operator (e.g., max) that depends on the specifications of the analysis task (details are provided in Section 6.2). Table 1 outlines the basic terms introduced throughout the description of the proposed framework, while their functional relations are demonstrated in Fig. 1.

Table 1. Legend of Introduced Terms

Term	**Symbol**	**Role**
Trained Classifier	F_c	- Degree of confidence that I_q depicts c
Domain Knowledge	$K_D = S(N_C, R, O)$	- Determine what evidence to look for. - Qualitatively relations between evidence and hypotheses.
Application Context	$X = S(app, W)$	- Determine where to look for evidence(i.e., application specific information, app). - Quantitative relations between evidence and hypotheses, W_{ij} (i.e., frequency of co-occurrence).
Hypotheses	$H(I_q) = \{Pr(c_1^H \mid I_q), \ldots, Pr(c_M^H \mid I_q)\}$	- Degrees of confidence for the concepts of c^H, as determined by $N_C \in K_D$ and $app \in X$, obtained by applying classifiers similar to F_c.
Evidence	$E(I_q) = \{Pr(c_1^E \mid I_q), \ldots Pr(c_K^E \mid I_q)\}$	- Degrees of confidence for the concepts of c^E, as determined by $N_C \in K_D$ and $app \in X$, obtained by applying classifiers similar to F_c.
Evidence driven probabilistic inference	$c = \arg \otimes_{c^H} (\acute{H}(I_q))$ where $\acute{H}(I_q) = Pr(c^H \mid H(I_q), R, O, W, E(I_q))$	- Perform inference by calculating $\acute{H}(I_q))$, using $E(I_q)$ as triggering evidence, $R, O \in K_D$ as belief propagation tracks and $W_{ij} \in X$ as causality quantification functions.

Based on these modules we aim to develop a decision support framework that derives directly from the knowledge structure, retains intact the inference tracks of logic, wraps probabilistically the causality links between domain concepts and handles uncertain estimations meaningfully. The ability of Bayes' theorem

to compute the posterior probability of a hypothesis by relating the conditional and prior probabilities of two random variables, was the reason for considering the use of bayesian networks for our purpose.

4 Bayesian Networks & Probabilistic Inference

Bayes' theorem can be used to update or revise beliefs in light of new evidence that are estimated with a certain amount of confidence. Adjusting this description to the formulation of Section 3, every time a classifier is applied on a visual representation, a hypothesis is formed around concept c and the visual representation I_q. The goal is to verify or reject the hypothesis stating that I_q depicts c, using the evidence $E(I_q)$. A bayesian network 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. Hence, a bayesian network can be used to facilitate three dimensions of perception: a) provide the means to store and utilize domain knowledge K_D, an operation that is served by the network structure and prior probabilities, b) organize and make accessible information coming from context $X \in S(app, W)$, which is supported by the Conditional Probability Tables (CPTs) attached to each network node and c) allow the propagation of evidence beliefs using message passing algorithms, an action facilitated by the Bayes' theorem. A methodology for consistently transforming ontologies into bayesian networks is essential for enabling evidence driven probabilistic inference. For the purposes of our work we adopted a variation of the methodology introduced in [7]. The proposed variation is mainly focused on the method employed for calculating the CPTs, as detailed later in this section.

Network Structure: Intuitively, deciding on the structure of a bayesian network based on an ontology can be seen as determining a function that maps ontological elements (i.e., concepts and relations) to graph elements (i.e., nodes and arcs). All translation rules described in [7] were implemented for determining the network structure out of an OWL ontology. The resulting network consists of concept nodes n_{cn} and control nodes n_{cl} (both of them having two states i.e., true and false) that are used to model the domain concepts and the associations between them, respectively. At this point, it is important to notice that the methodology described in [7] is only able to handle a limited set of constructors, namely owl:intersectionOf, owl:unionOf, owl:complementOf, owl:equivalentClass and owl:disjointWith, and as a consequence these are the constructors supported by our framework.

Parameter Learning: While the network structure encodes the qualitative characteristics of causality, (i.e., which nodes affect which), network parameters are used to quantify it, (i.e., how much is a node influenced by its neighbors). CPTs are used to capture the amount of this influence/impact and make it available for inferencing as part of the context structure $W_{ij} \in X$. The methodology adopted in this paper differs from [7] in what refers to the estimation of the network original probability distribution. While in [7] this information is provided explicitly by an expert, in our case it is learned from observation data, using the

Expectation Maximization (EM) algorithm [14]. More specifically, the prior and conditional probabilities for each concept node n_{cn} of the bayesian network, are initially calculated before considering any DL constructors. Subsequently, the DL constructors are migrated by inserting into the resulting network the appropriate control nodes n_{cl}. Once the structural translation has been completed, the CPTs for all concept nodes n_{cn} are re-calculated. Since no observation data are available for the control nodes n_{cl}, these nodes are treated as latent variables with two states (i.e., true and false). The last step is to specifically set the CPTs of all control nodes n_{cl} as appear in [7] and fix their states to "True", so as to enforce the semantic constraints expressed by the DL constructors.

Evidence-driven Probabilistic Inference: A framework that will allow beliefs to seamlessly flow over the established network is required. Pearl [15] introduced a message passing mechanism where messages are exchanged between father and child nodes carrying the information required to update their beliefs. In order to overcome the fact that Pearl's algorithm suffer from scalability issues, Lauritzen and Spiegelhalter [16] exploit a range of local representations for the network joint probability distribution, introducing the junction tree [17]. To the best of our knowledge, this is the most efficient and scalable belief propagation algorithm and will be the one used in our experiments.

5 Framework Functional Settings

5.1 Image Analysis Tasks

For carrying out an image analysis task using the proposed framework it is important to specify the following: a) formulate the hypothesis set $H(I_q)$ before initiating the decision mechanism, b) determine the methods used to obtain the initial confidence values of the evidence $E(I_q)$ and c) clarify what is considered to be the task specific analysis context $app \in X$, used to derive the evidence.

Image Categorization, involves selecting a category concept c_i describing the image as a whole. A hypothesis is formulated around each of the categories and with respect to the overall image, $H(I_q) = \{Pr(c_i|I_q) : \quad i = 1, \ldots, n\}$ where n is the number of category concepts. Global classifiers (i.e., models trained using image global information) are employed to estimate the initial likelihood for each hypothesis, $Pr(c_i|I_q)$. Regional concept information obtained by analyzing specific regions $I_q^{s_j}$ of the image at hand, is considered to be the source of contextual information $app \in X$ of this task. Local classifiers (i.e., models trained using image regional information) are applied on these regions and generate a set of confidence values that constitute the analysis 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 and m the number of identified regions. The distinction between the category concepts c_i (i.e., hypothesis concepts c^H in this case) and regional concepts $\acute{c}_i$ (i.e., evidence concepts c^E in this case) as well as their exact nature is determined by K_D.

Localized Image Region Labeling, annotates each of the identified regions with one of the available regional concepts $\acute{c}_i$. A hypothesis is formulated

for each of the available regional concepts and with respect to each of the regions identified in the image, $H(I_q) = \{Pr(\acute{c}_i|I_q^{s_j}) : \quad i = 1, \ldots, k \quad \& \quad j = 1, \ldots, m\}$ where k is the number of regional concepts and m is the number of identified regions. Regional classifiers are utilized to estimate the initial likelihood for each of the formulated hypotheses, $Pr(\acute{c}_i|I_q^{s_j})$ with $i = 1, \ldots, k$ & $j = 1, \ldots, m$. In this case, global image information is considered to be the source of contextual information $app \in X$ and the confidence values for each of the category concepts c_i, constitute the analysis evidence of this task, $E(I_q) = \{Pr(c_i|I_q) : \quad i = 1, \ldots, n\}$, where n is the number of category concepts. Once again, the knowledge structure K_D determines which concepts should be considered category concepts and which regional. However, since the nature of this task is different from image categorization, in this case $\acute{c}_i \equiv c^H$ and $c_i \equiv c^E$.

It is clear that the objective of our framework in both tasks is to operate on top of the classifiers' outcome with the aim to compensate for misleading decisions. Intuitively, the framework incorporates contextual information by favoring the co-occurrence of evidence that are known from experience to correlate. Additionally, the framework attempts also to exploit semantic restrictions, saying for instance that two concepts are disjointed. Therefore, provided that the majority of evidence coming from context are relatively strong and accurate, the framework is expected to make the correct decision by absorbing any misleading cues produced by the erroneous analysis of visual stimulus.

5.2 Low-level Image Processing

For low-level image processing we employed the scheme utilized in [8]. Four different visual descriptors proposed by the MPEG-7 standard [18] namely Scalable Color, Homogeneous Texture, Region Shape, Edge Histogram comprised the feature space. An extension of the Recursive Shortest Spanning Tree algorithm [19] was employed for producing a segmentation mask $S = \{s_i, \quad i = 1, \ldots, N\}$, with s_i representing the identified spatial regions. Finally, Support Vector Machines (SVMs) [20] as implemented by the libsvm library [21], were chosen to construct the statistically trained models, using the distance from the decision boundary in the kernel space as a way to measure the degree of confidence.

6 Experimental Study

The purpose of our experimental setup was to demonstrate the improvement in performance introduced by exploiting context and knowledge, compared to schemes that rely solely on low-level visual information. A dataset from the "Personal Collection" domain was selected for testing the proposed framework using the analysis tasks of Section 5.1.

6.1 Experimental Platform

Test set Characteristics: A collection I of 648 jpeg images comprised the test platform. Six different categories formulating the global (i.e., category) concepts

lexicon $C_G = \{Countryside\_buildings, Seaside, Rockyside, Forest, Tennis, Roadside\} \in N_C$, were used to manually characterize all 648 images at global level. Respectively, 25 more fine grained concepts constituting the local (i.e., regional) concepts lexicon $C_L = \{Building, Roof, Tree, Stone, Grass, Ground, Dried-plant, Trunk, Vegetation, Rock, Sky, Person, Boat, Sand, Sea, Wave, Road, Road-line, Car, Court, Court-line, Board, Gradin, Racket\} \in N_C$, were used to manually annotate the images at region level. A domain expert was employed to provide the logical relations between the elements of C_G and C_L using the OWL-DL ontology language, Fig. 2. The matching bayesian network automatically derived according to the methodology presented in Section 4 is depicted in Fig. 3.

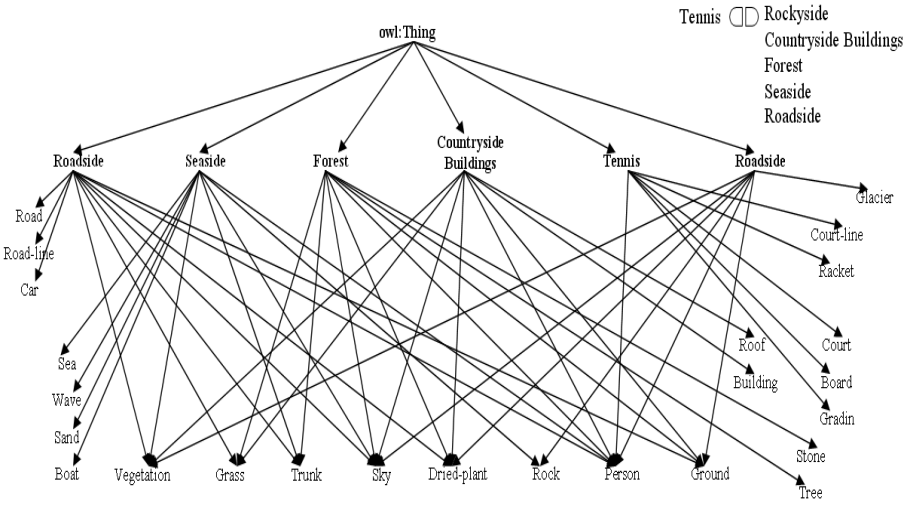


Fig. 2. Manually constructed ontology for describing the "Personal Collection" domain.

Calculating the prior probabilities and CPTs for each node of the constructed network requires a set of observation data. A subset of the manually annotated image set containing 324 samples was utilized to provide the parameter learning algorithm with the necessary observations and also to train the necessary classifiers, F_c, for each of the elements in C_G and C_L. Fig. 3 depicts the prior probabilities of all network nodes calculated by the learning algorithm. The remaining 324 images were used for testing.

6.2 Performance Evaluation

For measuring efficiency recall, precision and F-Measure were utilized. Based on the analysis tasks specified in Section 5.1, we have conducted the following experiments.

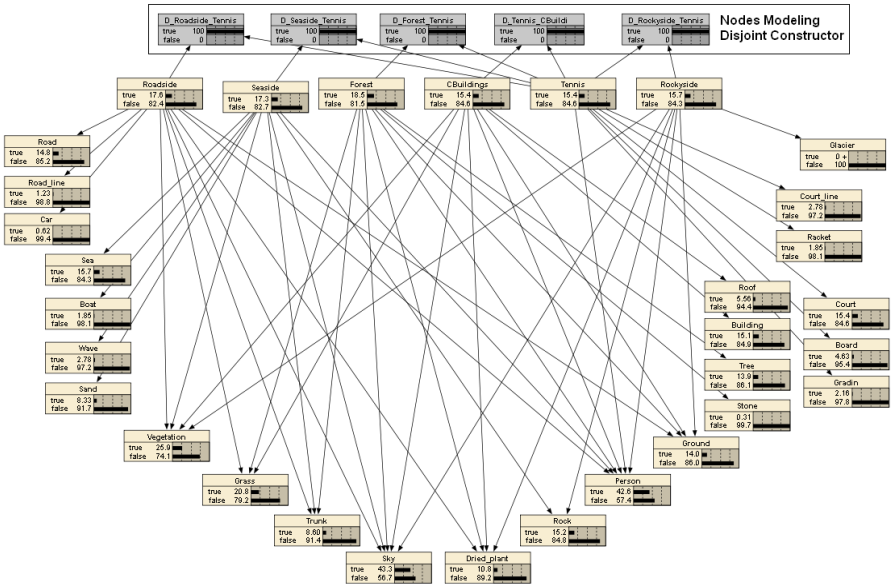


Fig. 3. Bayesian network automatically derived from the ontology of Fig. 2.

Image Categorization: In this experiment we measure the efficiency of categorizing all test images in one of the categories in C_G, using three configurations varying in the amount of utilized context and knowledge. In the first configuration we assess the performance of image categorization based solely on the output of global classifiers. In the second configuration, information coming from the local classifiers is incorporated into the network for helping towards the correction of the decisions erroneously taken by the global classifiers. In this case context and knowledge are utilized to extract the existing evidence and facilitate the process of evidence driven probabilistic inference. However, no semantic constraints (i.e., DL constructors) originating from the domain are incorporated into the decision process. This is the purpose of the last configuration where such constraints are incorporated into the bayesian network using the methodology presented in Section 4.

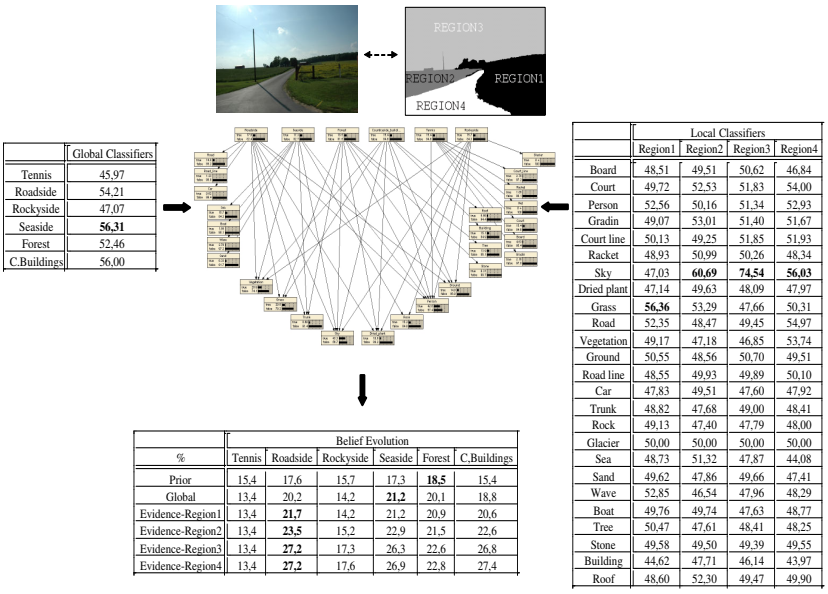
After formulating the hypotheses set for all category concepts, the framework looks for the presence of all regional concepts determined in K_D. All classifiers, global and local, are applied to formulate one set of confidence values for the image as a whole, $LK_{global} = \{Pr(c_i|I_q) : \quad \forall c_i \in C_G\}$ and one set per identified image region, $LK_{local} = \{Pr(c_j|I_q^{s_k}) : \quad \forall c_j \in C_L \quad \& \quad \forall s_k \in S\}$. All values of LK_{global} and the maximum per row values of LK_{local} are inserted as evidence into the bayesian network. Subsequently, the network is updated to propagate evidence impact and the category with the highest resulting likelihood is selected as the final decision (i.e., in this case $\otimes \equiv \max$). Table 2 summarizes the results for each of the framework configurations mentioned earlier.

Table 2. Image Categorization Evaluation Matrix

	%	Tennis	Roadside	Rockyside	Seaside	Forest	C. Build.	Avg
Global Classifiers only	Re	**100.00**	68.42	68.62	85.71	**76.66**	30.00	71.57
	Pr	83.33	**69.64**	70.00	67.60	63.88	**100.00**	75.74
	F-M	90.90	**69.02**	69.30	75.59	69.69	46.15	70.11
Global Classifiers Local Classifiers Know. & Context	Re	98.00	**73.68**	64.70	**91.07**	71.66	**54.00**	75.52
	Pr	90.74	64.61	**76.74**	**70.83**	**71.66**	90.00	77.43
	F-M	94.23	68.85	70.21	**79.68**	**71.66**	67.50	75.36
Global Classifiers Local Classifiers Know. & Context Sem. Constraints	Re	94,00	**73,68**	**70,58**	**91,07**	71,66	**56,00**	**76,17**
	Pr	**100,00**	64,61	76,59	69,86	70,49	90,32	**78,65**
	F-M	**96,90**	68,85	**73,46**	79,06	71,07	**69,13**	**76,41**

The performance achieved by the framework using the second configuration (row II of Table 2) is improved by $\approx 5\%$ (in terms of the F-Measure metric) compared to the first configuration (row I of Table 2). We will use the running example of Fig. 4 to demonstrate how evidence collected using regional information can revise a decision erroneously taken from a global classifier. By applying all global classifiers on the test image of Fig. 4 we get the probabilities of "*Global Classifiers*" table. According to these values the image should be characterized as *Seaside* since the corresponding classifier exhibit maximum confidence. The situation remains unaltered, as shown in the second row of "*Belief Evolution*" table, when the confidence values of all global classifiers are inserted into the network. However, this is not the case when the regional evidence i.e., the maximum value from each column of the "*Local Classifiers*" table are consecutively inserted into the bayesian network. The last four rows of "*Belief Evolution*" table illustrate how the likelihood of each category evolve in the light of new evidence. Eventually the correct category, *Roadside*, is found to exhibit maximum likelihood. What is interesting is the fact that only two out of four local classifiers (regions 1 and 3) succeeded in correctly predicting the depicted regional concept. Nevertheless, this information was sufficient for the evidence driven image analysis framework to infer the correct prediction, since the relation between the evidence *grass* identified in region 1 and the *Roadside* category, was strong enough to raise the inferred likelihood of this category above the corresponding value of *Seaside*, a category that receives no support by this evidence, as shown in Fig.2.

By examining the confusion matrix of TABLE 3 that corresponds to the second configuration of our framework, in conjunction with Fig. 2, where the amount of evidence shared between different image categories is depicted, it is clear that the system tends to confuse categories that share many visual characteristics. Another interesting observation derived from Fig. 2 concerns the small



	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	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. 4. An example of evidence-driven probabilistic inference for image categorization

Table 3. Confusion Matrix for Image Categorization (2^{nd} Configuration)

%	Tennis	Roadside	Rockyside	Seaside	Forest	C. Buildings
Tennis	98.00	0	0	2.00	0	0
Roadside	1.75	73.68	0	8.77	10.53	5.26
Rockyside	5.88	3.92	64.71	5.88	19.61	0
Seaside	0	5.36	3.57	91.07	0	0
Forest	0	10.00	8.33	10.00	71.67	0
C. Buildings	2.00	24.00	6.00	12.00	2.00	54.00

amount of evidence that *Tennis* shares with the rest of image categories. This is in fact a piece of information (i.e., semantic constraint) that originates from the domain and can be used to aid image analysis (i.e., third configuration of our framework). Indeed, by enhancing the ontology of Fig. 2 to associate *Tennis* with all other concepts in C_G with the *"owl:disjointWith"* DL-constructor, reconstructing the bayesian network to include the additional control nodes (see Fig. 3) and calculating the updated CPTs, the performance is further increased reaching an overall improvement of $\approx 6.5\%$ (row III of Table 3).

Region Labeling: The use of context and knowledge for region labeling was motivated by the fact that information concerning the image as a whole can potentially favor one region interpretation over another. Hence, it is clear

that the proposed framework can basically benefit region labeling when there is a conflict between the predictions suggested by global and local classifiers. If we let $Child(c_k : \quad c_k \in C_G) = (c_j : \quad \forall c_j \in C_L \quad where \quad k \rightarrow_{parent} j)$ be the subset of C_L corresponding to the child nodes of c_k according to K_D, $LK_{global} = \{Pr(c_i|I_q) : \quad \forall c_i \in C_G\}$ be the set of confidence values obtained from the global classifiers 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 on region s_w, 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 this case, since there is no reason to trust one suggestion over another, we make two hypotheses. The first assumes that the suggestion of the global classifier is correct and a global concept c_g is selected such as $g = \arg\max_i(LK_{global})$. Afterwards, the local concept c_l with maximum confidence that is included in the child node set of c_g is selected, such as $l = \arg\max_j(LK^{s_w}_{local})$ and $c_l \in Child(c_g)$. Both confidence values of c_g and c_l are inserted into the network as evidence and the overall impact on the likelihood of the hypothesis stating that the region under examination s_w, depicts c_l is measured. The second approach considers that the suggestion of the local classifier $c_{\acute{l}}$ is the correct, selected such as $\acute{l} = \arg\max_j(LK^{s_w}_{local})$. The confidence values of the global classifiers that correspond to the parent nodes of $c_{\acute{l}}$ are examined and the one $c_{\acute{g}}$ with maximum value is selected, such as $\acute{g} = \arg\max_i(LK_{global})$ and $c_{\acute{g}} \in Parent(c_{\acute{l}})$. As in the previous case both likelihoods are inserted into the network and the overall impact on the likelihood of the hypothesis stating that the examined region s_w, depicts $c_{\acute{l}}$ is measured. Eventually, these values are compared and the concept corresponding to the largest value is chosen (i.e., this is the functionality of $\otimes$ operator for this case). If no conflict occurs the concept corresponding to the local classifier with maximum confidence is selected. Fig. 5 presents the evaluation results and shows that an average increase of approximately 4.5% is accomplished when the proposed framework is used. Regional concepts that exhibit zero hits from the local classifiers (i.e., Racket, Road line, Car, Glacier, Stone) are not included in the evaluation results.

7 Conclusions & Future Work

The problem of using visual evidence to assist image analysis has been thoroughly treated and a concrete framework addressing the identified issues has been proposed. The suitability of ontologies and bayesian networks for imitating some of the fundamental aspects of visual perception has been investigated. Experiments demonstrated that the proposed framework is able to analyze images using different configurations, in terms of the amount of utilized context and knowledge, and manage to achieve statistically significant improvement with respect to the solutions relying solely on visual stimulus.

One important prerequisite for allowing the proposed framework to maximize the performance gain, is to operate on a sufficiently large amount of training data. This is hindered by the fact that it is really a cumbersome procedure to manually annotate a sufficiently large number of images, especially at region

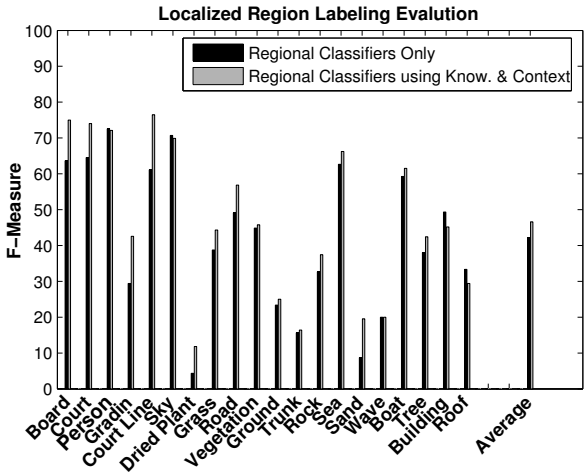


Fig. 5. Evaluation results for localized region labeling.

level, and in this way ensure that the sample data tendencies converge to true domain statistics. A solution to this problem could be to adjust the underlying image processing module so as to allow handling of large multimedia corpora that are being populated automatically, as in the case of WWW and Internet users. Given the fact that social sites like Flickr[1], accommodate image corpora that are being populated with hundreds of user tagged images on a daily basis and taking into consideration that literature has already reported efforts on performing localized region-label association, from weakly annotated data [22], pipelining such schemes with the proposed framework may help overcoming some of the problems deriving from the use of limited size training sets.

Acknowledgment This work was funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978.

References

1. Naphade, M.R., Kristjansson, T.T., Frey, B.J., Huang, T.S.: Probabalistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In: ICIP (3). (1998) 536–540
2. Naphade, M.R., Huang, T.S.: A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Transactions on Multimedia **3**(1) (2001) 141–151

[1] http://www.flickr.com/

3. Luo, J., Savakis, A.E., Singhal, A.: A bayesian network-based framework for semantic image understanding. Pattern Recognition **38**(6) (2005) 919–934
4. Luo, J., Savakis, A.E.: Indoor vs outdoor classification of consumer photographs using low-level and semantic features. In: ICIP (2). (2001) 745–748
5. Kane, M.J., Savakis, A.E.: Bayesian network structure learning and inference in indoor vs. outdoor image classification. In: ICPR (2). (2004) 479–482
6. Matos, L.N., de Carvalho, J.M.: Combining global and local classifiers with bayesian network. In: ICPR (3). (2006) 1212–1215
7. Ding, Z., Peng, Y., Pan, R.: A bayesian approach to uncertainty modeling in owl ontology. In: Int. Conf. on Advances in Intelligent Systems - Theory and Applications. (November 2004)
8. Papadopoulos, G.T., Mezaris, V., Kompatsiaris, I., Strintzis, M.G.: Combining global and local information for knowledge-assisted image analysis and classification. EURASIP J. Adv. Signal Process **2007**(2)
9. Athanasiadis, T., Mylonas, P., Avrithis, Y., Kollias, S.: Semantic image segmentation and object labeling. IEEE Transactions on Circuits and Systems for Video Technology **17**(3) (March 2007) 298–312
10. Town, C.: Ontological inference for image and video analysis. Machine Vision and Applications **17**(2) (2006) 94–115
11. Cardoso, J.: The semantic web vision: Where are we? IEEE Intelligent Systems **22**(5) (2007) 84–88
12. McGuinness, D.L., van Harmelen, F.: OWL web ontology language overview. W3C recommendation, W3C (February 2004) http://www.w3.org/TR/2004/REC-owl-features-20040210/.
13. Horrocks, I.: Description logics in ontology applications. In: Automated Reasoning with Analytic Tableaux and Related Methods. (2005) 2–13
14. McLachlan, G.J., Krishnan, T.: The EM algorithm and extensions. 2nd edn. John Wiley and Sons (1997)
15. Pearl, J.: Fusion, propagation, and structuring in belief networks. Artif. Intell. **29**(3) (1986) 241–288
16. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. (1990) 415–448
17. Jensen, F.V., Jensen, F.: Optimal junction trees. In Kaufmann, C.M., ed.: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, San Mateo (1994)
18. Manjunath, B.S., Ohm, J.R., Vinod, V.V., Yamada, A.: Colour and texture descriptors. IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7 **11**(6) (Jun 2001) 703–715
19. T. Adamek, N. O'Connor, N.M.: Region-based segmentation of images using syntactic visual features. In: WIAMIS'05, Montreux, Switzerland (2005)
20. Scholkopf, B., Smola, A., Williamson, R., Bartlett, P.: New support vector algorithms. Neural Networks **22** (2000) 1083–1121
21. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines (2001)
22. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: ECCV (4). (2002) 97–112