Image interpretation by combining ontologies and bayesian networks

S. Nikolopoulos[1,2], G. Th. Papadopoulos[1], I. Kompatsiaris[1], and I. Patras[2]

[1] CERTH-ITI, Informatics and Telematics Institute, Greece, {nikolopo@iti.gr, papad@iti.gr, ikom@iti.gr}
[2] School of Electronic Engineering and Computer Science, QMUL, UK {i.patras@eecs.qmul.ac.uk}

Abstract. A drawback of current computer vision techniques is that, in contrast to human perception that makes use of logic-based rules, they fail to benefit from knowledge that is provided explicitly. In this work we propose a framework that performs knowledge-assisted analysis of visual content using ontologies to model domain knowledge and conditional probabilities to model the application context. A bayesian network (BN) is used for integrating statistical and explicit knowledge and perform hypothesis testing using evidence-driven probabilistic inference. Our results show significant improvements compared to a baseline approach that does not make any use of context or domain knowledge.

1 Introduction

The advances in information technology have reduced the spatial and temporal obstacles in information exchange, allowing users to easily generate and exchange large amounts of digital data. However, the limitations of machine understanding makes it difficult for automated systems to interpret and index all this content in a manner coherent with human cognition. 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.

In our work we introduce a framework for enhancing image analysis using different types of evidence. As evidence we define the information that can be used to support or disproof a hypothesis. In our framework (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. The application context and the domain knowledge are considered to be the a priori/fixed information, while the visual stimulus depends on the examined image and is considered to be the observed/dynamic information. We model the layer of evidence so as to effectively combine both a priori and observed information. More specifically, first we 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. The main contributions of our work are: 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. b) We show how global and regional evidence can be probabilistically combined within a BN that incorporates domain knowledge and application context.

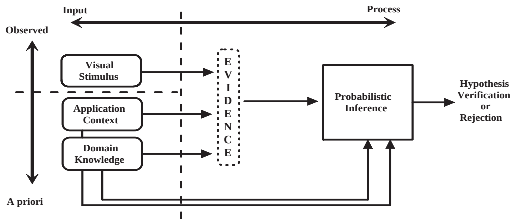


Fig. 1. Functional relations between the different components of our framework.

2 Related Work

Semantic image analysis has been addressed by mapping low-level visual features (i.e., color, shape) to high-level descriptions (i.e., concepts), without using domain knowledge or context. Some indicative works include [1] where the authors use the mean of global image features to represent the gist of a scene, and [2] where scene classification is performed using bayesian classifiers. However, the suboptimal performance of these solutions has motivated the exploitation of knowledge and context.

Towards this objective the authors of [3] introduce “Multijects” as a way to map time sequence of multi-modal, low-level features to higher level semantics and “Multinets” for representing higher-level probabilistic dependencies between “Mutlijects”. In the same lines, [4] proposes a framework for semantic image understanding that integrates in the same knowledge-based inference framework (based on BNs), both low-level and semantic features. Similarly, [5] uses low-level features and a BN to perform indoor versus outdoor scene categorization. However, the absence of a methodology for integrating domain knowledge into the inference process is what differentiates these works from our approach. Finally there are also works that utilize ontologies as a means to encode domain knowledge. [6] presents a method for combining ontologies and BNs in an effort to introduce uncertainty in ontology reasoning and mapping, while [7] proposes a knowledge assisted image analysis scheme that combines local and global information. However, none of these works attempt to couple ontology-based approaches with probabilistic inference algorithms for combining concept detectors, context and knowledge.

3 Framework Description

Visual Stimulus: For analyzing the visual stimulus we employ supervised learning where a classifier is trained to identify a concept, provided that a sufficiently large number of examples are available. If N_C denotes the set of domain concepts, a concept detector can be implemented using a classifier F_c that is trained to recognize instances of the concept $c \in N_C$. If 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.
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. We use OWL–DL to construct a structure $K_D = S(N_C, R, O)$ that describes how the domain concepts are related to each other. DL stands for "Description Logics" [8] and constitutes a specific set of constructors such as intersection, union, disjoint, complement, etc. 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.
Application Context: Let *app* denote the application specific information used to guide the analysis mechanism in searching for evidence, and $W = [W_{i,j}]$ the matrix whose elements quantifies the effect of concept c_i on c_j. Then, we consider the application context $X = S(app, W)$ to consists of both *app* and W. W_{ij} is implicitly extracted from data and encoded into the Conditional Probability Tables (CPTs) of the BN to influence the probabilistic inference process.
Evidence-driven Probabilistic Inference: To perform inference: a) we use K_D to decide which of the concepts should be treated as evidence c^E, b) we use *app* to decide where to physically search for them, c) we apply 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 concepts should constitute the hypotheses set c^H, e) we provide as soft evidence the confidence degrees for the concepts in c^E and trigger probabilistic inference in the BN, f) we propagate evidence beliefs using the network's inference tracks R and the causality quantification functions W_{ij}, and g) we calculate the posterior probabilities for all concepts in c^H. If $\acute{h}(I_q, c_i)$ are the posterior probabilities of the network nodes and $\otimes$ is an operator (e.g., max) that depends on the specifications of the analysis task, semantic image interpretation is achieved based on the formula: $c = \arg \otimes_{c_i \in c^H} (\acute{h}(I_q, c_i))$.

4 Ontology to Bayesian Network mapping

Our motive for using BNs is to estimate the posterior probabilities of the concepts in the hypothesis set c^H, using the observed confidence degrees of the concepts in the evidence set c^E. The work in [6] 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.

Network Structure: The transformation of an ontology to a BN takes place in two stages. In the first stage, the BN incorporates the hierarchical information of the ontology by transforming all concepts into nodes (called concept nodes n_{cn}) with two states (i.e., true and false). 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 superclass-to-subclass direction. At the second stage, the BN incorporates the semantic constraints of the ontology by creating a control node n_{cl} for each DL constructor (see [6] for details). The constructors that can be handled are owl:intersectionOf, owl:unionOf, owl:complementOf, owl:equivalentClass and owl:disjointWith.

Parameter Learning: Once the structure is fixed, each concept node n_{cn} needs to be assigned a prior probability if it is a root node or a CPT if it is a child node. In [6] these probabilities are set by domain experts. 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 this approach where the necessary probabilities are learned from data (i.e., concept label annotations of the images). The conditional probabilities of all concept nodes are learned by employing the Expectation Maximization (EM) algorithm on sample data. The last step is to manually set the CPTs of all control nodes n_{cl} as shown in [6] 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.

5 Framework functional settings

Our framework implements two different image analysis tasks: (a) **Image categorization** selects 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 $h(I_q)$. Global classifiers 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 local information (e.g., knowing that a region depicts *road* is a piece of contextual information that can help deciding whether the image depicts a *Seaside* or a *Roadside* scene). Local classifiers are applied to the pre-segmented regions $I_q^{s_j}$, in order to generate the pieces of evidence $E(I_q)$ that will be used to trigger probabilistic inference. (b) **Localized region labeling**, assigns 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. Local classifiers are used to estimate the initial probability for each of these hypotheses. Here, the contextual information *app* is considered to be the image as a whole (e.g., knowing that an image depicts a *Roadside* scene can help in deciding whether a specific region depicts *sea* or *road*). The confidence degrees of the category concepts c_i constitute the pieces of evidence for this task $E(I_q)$, which are used to trigger probabilistic inference. 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 by favoring the hypotheses with maximum positive impact on its posterior probability.

The low level processing of visual stimulus consists of visual features extraction, segmentation and learning the concept detection models. Four MPEG-7 visual descriptors [9], namely Scalable Color, Homogeneous Texture, Region Shape, and Edge Histogram, were employed as described in [7]. Segmentation was performed using an extension of the Recursive Shortest Spanning Tree algorithm [10] and Support Vector Machines (SVMs) with a gaussian radial kernel function were employed for learning the concept detection models.

6 Experimental Study

In our study we demonstrate the performance improvements achieved by exploiting context and knowledge compared to baseline detectors that rely solely on visual information. A collection of 648 annotated at global and region detail comprised our dataset[3]. Half of the images were used for training the classifiers F_c and learning the BN parameters and the other half for testing. The resulting BN is depicted in Fig. 2.

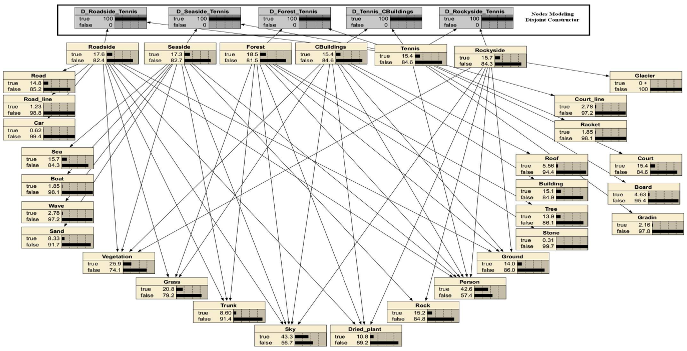

D_Roadside_Tennis true 100 false 0
D_Seaside_Tennis true 100 false 0
D_Forest_Tennis true 100 false 0
D_Tennis_CBuildings true 100 false 0
D_Rockyside_Tennis true 100 false 0
Nodes Modeling
Disjoint Constructor
Roadside true 17.6 false 82.4
Seaside true 17.3 false 82.7
Forest true 18.5 false 81.5
CBuildings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

Fig. 2. The nodes in the black frame are used to model the disjointness between the *Tennis* and all other category concepts in the domain.

Image categorization is evaluated using three configurations. In the baseline configuration $CON1$ we assess the performance of image categorization based solely on visual stimulus. The second configuration $CON2$ uses context and knowledge in order to extract the existing evidence and facilitate the process of evidence driven probabilistic inference. The BN employed in this configuration is the one depicted in Fig. 2 without the nodes enclosed by the black frame. The third configuration $CON3$ takes into account the semantic constraints of the domain. In this case, the utilized BN is extended with the addition of the

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

control nodes (i.e., the nodes enclosed by the black frame of Fig. 2) that are used for modeling the disjointness between *Tennis* and all other category concepts. The reason for treating $CON2$ and $CON3$ as two different configurations was to examine how much of the improvement comes from the use of regional evidence and concept hierarchy information ($CON2$), and how much comes from the enforcement of the semantic constraints ($CON3$).

In both $CON2$ and $CON3$ the analysis process unfolds as follows. Initially, we formulate the hypotheses set using all category concepts. Then, we search for 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, $LK_{global} = \{Pr(c_i|I_q) : \forall c_i \in C_G\}$ and one set per identified image region, $LK_{local} = \{Pr(c_j|I_q^{s_k}) : \forall c_j \in C_L \quad \& \quad \forall s_k \in S\}$. All values of LK_{global} and the maximum per column values of LK_{local} are introduced as soft evidence into the BN nodes. Then, the network is updated to propagate evidence impact and the concept corresponding to the node with the highest resulting posterior probability (among the category concepts), is selected to categorize the image (i.e., in this case $\otimes \equiv$ max, see Section 3). Fig. 3(a) shows that CON2 outperforms CON1 by $\approx 5\%$ on average. The running example of Fig. 4 demonstrates how evidence collected using regional information ($CON2$) can correct a decision erroneously taken by a global classifier that relies solely on visual stimulus ($CON1$). Finally, using CON3 the performance is further increased with an average improvement of $\approx 6.5\%$, compared to the baseline ($CON1$). Given that the semantic constraint was enforced between the *Tennis* and all other concepts in C_G, the improvement in performance comes from the correction of the test samples that were originally mis-categorized as *Tennis*.

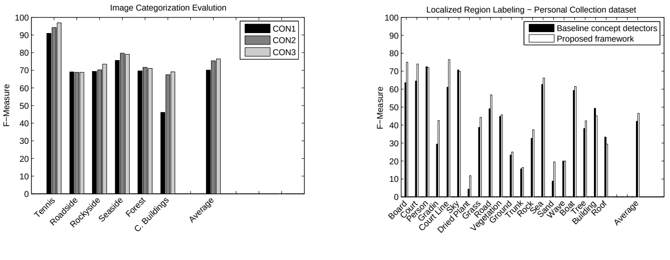
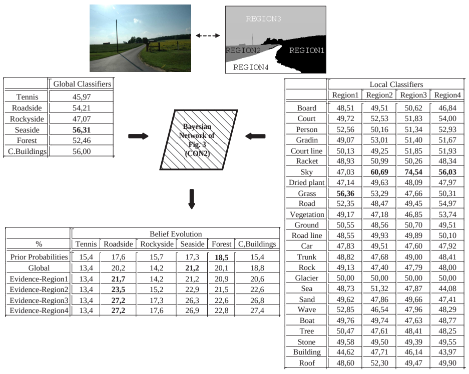


Fig. 3. a) F-Measure scores for image categorization using CON1,CON2 and CON3 configurations, and b) F-Measure scores for localized region labeling.

Localized Region Labeling was performed using the BN of Fig. 2 (without the nodes enclosed by the black frame). Our framework is put into force when there is a conflict between the decisions suggested by the global and local classi-



	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. 4. 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.

fiers. 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 global confidence values for 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 local confidence values for 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 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 that $I^{s_w}_q$ 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. 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 that $I^{s_w}_q$ depicts $c_{\acute{l}}$ is measured. Eventually, the values of the two different cases are compared and depending on the largest, c_l or $c_{\acute{l}}$ is chosen to label the region in question (i.e., this is the functionality of $\otimes$ operator described in Section 3, for this task). If no conflict occurs, the concept corresponding to the local classifier with maximum confidence is selected. Fig. 3(b) shows that when using the proposed framework an average increase of approximately 4.5% is accomplished. Finally, Table 1 shows how our method compares with two state-of-the art methods [11] and [12] on the MSRC dataset[4].

[4] http://research.microsoft.com/vision/cambridge/recognition

Table 1. 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 [11]	**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 [12]	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

7 Conclusions

Our experiments have shown that the amount and nature of the semantic information that can be used to enhance image interpretation depends on the characteristics of the domain. Although the knowledge structure and the causality relations were useful in all cases, the semantic constraints originating from the domain were only able to help when the imposed rules were sufficiently concrete (e.g., the disjointness between "Tennis" and all other category concepts). On the contrary, attempts to incorporate semantic constraints that were less strict from the visual inference point of view didn't lead to performance improvements.

References

1. 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.
2. D. Gokalp and S. Aksoy, "Scene classification using bag-of-regions representations," in *CVPR '07*, june 2007, pp. 1 –8.
3. M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE TMM*, vol. 3, no. 1, pp. 141–151, 2001.
4. J. Luo, A. E. Savakis, and A. Singhal, "A bayesian network-based framework for semantic image understanding," *Pattern Recognition*, vol. 38, no. 6, 2005.
5. 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.
6. Z. Ding, Y. Peng, and R. Pan, "A bayesian approach to uncertainty modeling in owl ontology," in *Int. Conf. on Adv. in Intel. Sys. - Theory and Applications*, 2004.
7. 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.
8. I. Horrocks, "Description logics in ontology applications," in *Automated Reasoning with Analytic Tableaux and Related Methods*, 2005, pp. 2–13.
9. B. S. Manjunath, J. R. Ohm, V. V. Vinod, and A. Yamada, "Colour and texture descriptors," *IEEE TCSVT, Special Issue on MPEG-7*, vol. 11, pp. 703–715, 2001.
10. N. M. T. Adamek, N. O'Connor, "Region-based segmentation of images using syntactic visual features," in *WIAMIS '05*, Montreux, Switzerland, 2005.
11. 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.
12. J. Verbeek and B. Triggs, "Region classification with markov field aspect models," *CVPR '07*, pp. 1–8, 2007.