Bayesian Networks Seminar Paper: Applying Bayesian Networks to Semantic Image Understanding

Xin Guo†, Yuanmin Xu*
†*s1882201* **s1929313*
† **Leiden Institute of Advanced Computer Science, Leiden University*
Email: †{guoxin1122}@hotmail.com,{xymhappy6}@gmail.com*

***Abstract*—Computer vision is an interdisciplinary field that deals with how computers can be made for gaining high-level understanding from digital images or videos. The techniques of the computer vision have made considerable progress in recognizing object categories. Current research in content-based semantic image understanding is largely confined to exemplar-based approaches built on low-level feature extraction and classification. In this paper, we introduce two relative papers which subject in the computer vision and image interpretation using the Bayesian Network. Two papers are presented to deal with semantic image understanding problem using the Bayesian networks. The first paper, published in 2005, presents a general-purpose knowledge integration framework that employs BN in integrating both low-level and semantic features, and applies this framework to detecting main photographic subjects. The second paper, published in 2011, proposes a more powerful framework which can identify distinct scenes in the image using evidence-driven probabilistic inference. We study the methods they proposed as well as the cases and we conclude and make a discussion eventually.**

***Keywords*-Computer Vision; Bayesian Network, Image Interpretation; Classifier; Ontology**

I. Introduction

Computer vision is an interdisciplinary field that deals with how computers can be made for gaining high-level understanding from digital images or videos. It refers to using the cameras and computers instead of human eyes to identify, track and measure the targets and further to do image processing the images to the ones that are more suitable for humans. The tasks of the computer vision include methods for acquiring, processing, analyzing and understanding digital images. Digital images are used everywhere and cares about our daily life. For example, major uses of imaging based on gamma rays include nuclear medicine and astronomical observations. In nuclear medicine, the approach is to inject a patient with a radioactive isotope that emits gamma rays as it decays. Images are produced from the emissions collected by gamma ray detectors.

A classifier is an algorithm that takes a set offeatures that characterize objects and uses them to determine the class of each object. In supervised classification, meaning that a human expert both has determined into what classes an object may be categorized and also has provided a set of sample objects with known classes. This set of known objects is called the training set because it is used by the classification programs to learn how to classify objects. Image interpretation acts of examining images to identify objects and judge their significance. Interpreting images in terms of their semantic content has primarily been addressed by devising methods that map low-level image visual characteristics to high-level descriptions without making any use of domain knowledge and application context. Image interpretation comprises at least three mental acts that may or may not be performed simultaneously: The measure of images of objects; identification of the objects imaged and appropriate use of this information in the solution of the problem. Therefore, the use of classification and image interpretation is import and the basis of the computer vision.

There are many methods to do the image interpretation, such as decision-theoretic methods including the matching, optimum statistical classifiers, neural networks and etc and the structural methods such as matching shape numbers and string matching. Today we will focus on the classifier which is rooted in the Bayesian network (BN). BN is a probabilisticgraphical model and represents a set ofrandom variablesand theirconditional dependenciesvia directed acyclic graph (DAG). Each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods.

BN is widely used to the classification and image interpretation and it has proven to be an effective knowledge representation and inference engine in artificial intelligence and expert systems research. In [1], a BN can be utilized as an inference mechanism for facilitating a classification method based on feature space segmentation. The work presented in [2] presents a method for combining ontologies and BNs 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.

We have read two papers: A Bayesian network-based framework for semantic image understanding and Evidence-

Driven Image Interpretation by Combining Implicit and Explicit Knowledge in a Bayesian Network. Two papers are presented to deal with semantic image understanding problem using the Bayesian networks. The first paper, published in 2005, presents a general-purpose knowledge integration framework that employs BN in integrating both low-level and semantic features, and applies this framework to detecting main photographic subjects. The second paper, published in 2011, proposes a more powerful framework which can identify distinct scenes in the image using evidence-driven probabilistic inference.

The rest of the paper is organized as follows. In the next section, we give an overview of the first paper, also with the detailed introductions of the method of the classification. The advanced method introduced in the second paper is shown in III. Comparisons between two papers and our general ideas are discussed in section IV.

II. A Bayesian network-based framework for semantic image understanding

This paper focuses on the fusion technique of low-level feature and semantic features for scene interpretation. Here, the BN is used to explicitly integrate domain knowledge and to reduce a joint probability distribution to conditional independence relationships. They developed a fractional frequency counting-based training method to address the problem of partially certain ground truth, and a probabilistic reasoning approach to detecting main subject in the image. Finally, they describe a benchmarking study performed to compare the results of the BN-based automatic Main Subject Detection (MSD) system with other classifier systems.

A. Method

1) Framework: Fig. 1 illustrates the proposed general framework for semantic understanding of pictorial images. The input is a digital image of a natural scene. Two sets of descriptors are extracted from the image: the first set corresponds to low- level features, such as color, texture, and edges; the second set corresponds to semantic objects that can be automatically detected. The low-level features can be extracted on a pixel or block basis, using a bank of pre-determined filters aimed at extracting color, texture or edge characteristics from the image. The semantic features are obtained using a bank of pre-designed object detectors that have reasonable accuracy (e.g., at least better than chance). A Bayes net here consists of four components: (i) Priors: the initial beliefs about various nodes in the Bayes net; (ii) Conditional Probability Matrices (CPMs): knowledge about the relationship between two connected nodes in the Bayes net; (iii) Evidences: observations from feature detectors that are input to the Bayes net; (iv) Posteriors: the final computed beliefs after the evidences have been propagated through the Bayes net.

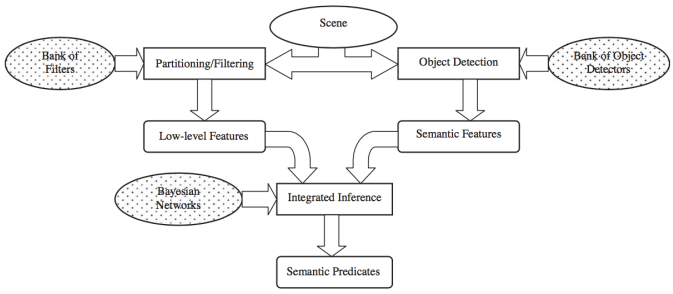

Figure 1: A BN for main subject detection

$$P(S|E) = \frac{P(S,E)}{P(E)} = \frac{P(E|S)P(S)}{P(E)}, \tag{1}$$

where S denotes semantic task and E denotes evidence. Probabilistic reasoning uses the joint probability distribution of a given domain to answer a question about this domain.

2) Training the Bayesian network parameters : They developed a fractional frequency counting-based training method. Fractional frequency counting-based training is very similar to the frequency-counting approach except that they now weight the feature measurements us- ing the ground truth. Thus, each feature measurement can now contribute towards all the labels of the parent node depending upon the ground truth associated with the parent node. Similarly, they allow the feature detector to provide partially certain evidences about the various labels associated with the child node. Thus, each complete training sample in this method contributes not just to one cell of the CPM, but, potentially, to all the cells. The CPM can be computed using fractional frequency counting as follows:

$$CPM = [(\sum_{i\in I}\sum_{r\in R_i} n_i F_r^T T_r)C]^T, \tag{2}$$

$$F_r = [f_0^r f_1^r f_2 \cdots f_m^r], T_r = [t_0^r t_1^r t_2 \cdots t_l^r], \tag{3}$$

$$C = diag p_j, p_j, p_j = (\sum_{i\in I}\sum_{r\in R_i} n_i t_r. \tag{4}$$

where I is the set of all training images, R_i is the set of all regions in image i, and n_i is the number of observations (observers) for image i. Moreover, F_r represents the m-label feature-evidence vector for region r, T_r represents the l-value ground-truth vector, and C denotes an $l \times l$ diagonal matrix of normalization constant factors.

The ground truth is now certain rather than probabilistic, since the main subject decisions made by each observer are binary. The frequency f would be expressed as:

$$f = \frac{\sum_{o\in O}\sum_{r\in R} T_{o,r} e_r}{\sum_{o\in O}\sum_{r\in R}}, \tag{5}$$

where O is the set of observers, R is the set of all regions, $T_{o,r}$ represents the ground truth value for region r from

observer o, and e_r represents the feature detector output for region r. Assuming there are N observers (N = above equation is equivalent to

$$f = \frac{\sum_{r \in R} \sum_{o \in O} T_{o,r} e_r}{N \sum_{r \in R}} \tag{6}$$

which is equivalent to

$$f = \frac{\sum_{r \in R} (\sum_{o \in O} T_{o,r}/N) e_r}{\sum_{r \in R}} \tag{7}$$

Once the BN has been constructed and trained, it can be used to compute the joint probability distributions very efficiently. The next section describes the use of the BN-based feature integration framework for an applications in the photographic image understanding domain.

B. Application to Main Subject Detection (MSD)

They developed a probabilistic reasoning approach to MSD. In particular, the algorithm consists of region segmentation, perceptual grouping, feature extraction, and probabilistic reasoning. First, an input image is segmented into a few regions of homogeneous (color) properties. Next, the region segments are grouped into larger regions corresponding to perceptually coherent objects with similar properties using non-object-specific grouping. These regions are evaluated for their saliency in terms of two independent, but complementary, types of features' structural and semantic. For example, recognition of human skin or faces is semantic while determination of what stands out generically is categorized as structural. For structural features, a set of low-level vision features (including color and texture) and geometric features is extracted. Semantic features can be further used to perform object-specific grouping which attempts to segment whole objects such as people or building in the image.

To integrate those diverse features, a multi-layer Bayes net is used to express the relationships between various feature detectors and its structure is designed based on domain knowledge [3] as shown in Fig.2, ensuring the conditional independence among various features. After evidence propagation through the entire network, the root node MainSubject gives the posterior belief that a region is part of the main subject. This node has two labels, MainSubject and Background. Since this is the root node, there is an a priori belief associated with its label set. Using data from training images and frequency counting, it was computed that the a priori belief is $P(MainSubject) = 0.28$ and $P(Background) = 0.72$. Examples of the experimental results are shown in Fig.3. The results are very encouraging in that most of the regions that belong to the main subject are differentiated from the background clutter in the image.

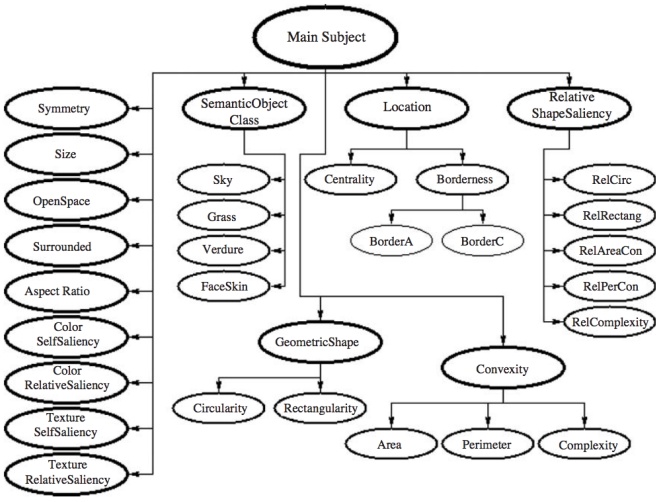

Figure 2: Communication mechanism

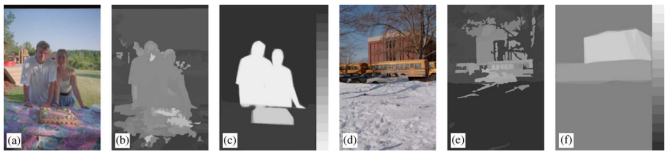

Figure 3: Examples of MSD results: (a,d) images; (b,e) MSD results; (c,f) ground truth maps

C. Benchmarking Bayes Net Performance

In this section, they describe a benchmarking study performed to compare the results of the BN-based automatic MSD system with other versions of the system. They implemented two Bayesian network classifiers, i.e. SLBN (single-level)and MLBN (multi-level), a nave central zone predictor, and two neural network based classifiers, one using a separate training and testing set, and the other using leave-one-out training.

Table I shows the number of images on which the MLBN performs better than the other classification schemes. The multilevel BN classifier beats the central zone predictor (Czone) and the SLBN classifier on both the train and the test image sets by ratios of approximately 2:1. It is similar to the training set neural network (NN-TS) overall, but does worse on the train image set and better on the test image set by similar ratios. It performs slightly worse than the leave-one-out neural network (NN-LOO) on both the image sets by similar ratios of approximately 3:4. More interestingly, the performance of the MLBN and the neural networks is comparable. This is also expected, as both the systems are able to use the full set of features and have similar expressive power. The true advantage of the BN lies not necessarily in increased performance gains (this would actually be hard since neural network and BN are theoretically equivalent), but in increased generalizability and ease-of-use. Unlike neural networks, the BN is extremely stable in the presence of missing or faulty feature detectors.

Fig. 4 shows the results of the null hypothesis tests performed on the dKS results for each ordered pair of classifiers. The tests were designed to check whether the performance of each classifier was statistically significantly better than the performance of the other four classifiers on the train and the test set of images. The table reads horizontally in that each row of the table tests for that classifier being statistically significantly better than the others. The MLBN-based classifier performs statistically significantly better than the central zone predictor and the SLBN-based classifier on both the train and the testing set of images.

	Training Set					Testing Set				
	CZone	SLBN	MLBN	NN TS	NN LOO	CZone	SLBN	MLBN	NN TS	NN LOO
CZone		F	F	F	F		F	F	F	F
SLBN	T .0044		F	F	F	F		F	F	F
MLBN	T .0006	T .0374		F	F	T .0091	T .0322		T .0016	F
NN(TS)	T .0004	T .0015	F		F	F	F	F		F
NN(LOO)	T .0005	T .0277	F	F		T .0284	F	F	T .0007	

Figure 4: Null hypothesis tests on dKS results for each pair of classifiers.

The MLBN also produces statistically significantly better results than the training set neural network on the testing set of images, although there is no statistically significant difference in their performance on the training set of images. This is to be expected as the training set neural network memorizes the training data to a certain degree and can reproduce those results fairly well. Increasing the size of the training set and imposing additional constraints on the neural network training method (such as a validation stop) can mitigate the memorization effect but will result in reduced performance from the neural network on the training set of images. There is no statistically significant difference between the performance of the multilevel BN and the leave-one-out neural network at the specified confidence level (5% error rate) on either of the two sets of images.

Fig. 5 presents the results of the analysis of variance tests on the train and the testing set of images. The analysis shows that in the case of the training set of images, the central zone predictor is statistically significantly worse than the remaining four classifiers. Also, the SLBN performs statistically significantly worse than the training set neural network on the train image set. There are no statistically significant differences between any of the remaining sets of classifiers. On the testing set, the central zone predictor and the training set neural network perform statistically significantly worse than the MLBN. Also, the leave-one-out neural network performs statistically significantly better than the central zone predictor. There are no statistically significant differences between any of the remaining sets of classifiers. As previously discussed, the main conclusions to be drawn from the benchmarking study are:

1 Using a set of features and a good inference algorithm (BN or neural network) leads to statistically significantly better performance than a naive predictor such as central zone.
2 The BN structure needs to be carefully constructed to account for dependencies between variables in the do- main. It also needs to be expressive (multi-level instead of single-level) to fully utilize the entire gamut of features available for the best performance.
3 BN are theoretically equivalent to neural networks and should result in similar performance when trained correctly. The primary advantage of the BN-based system comes from the flexibility, interpretability, and ease-of-use.

Table I: Performance of the multi-level Bayes network based classifier vs. other classifiers using the dKS metric

MLBN	Image set	CZone	SLBN	NN(TS)	NN(LOO)
# of images w/better performance	Train	35	36	16	17
	Test	33	25	30	20
	All	68	61	46	37
# of images w/worse performance	Train	13	10	27	23
	Test	15	16	15	25
	All	28	26	42	48

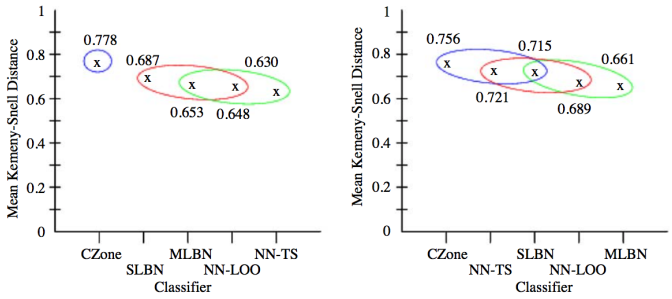

Figure 5: Analysis of variance test on the dKS results of the five classifiers: Left-raining set, Right-testing set.

D. Conclusion

In this paper, they presented a unified image understanding framework based on BN, where both low-level and semantic features can be incorporated for improved performance. In all three of the applications discussed in this paper (two of them are presented in this seminar paper), they demonstrated that the BN-based systems have excellent generalization on novel datasets. They attribute this to the fact that the training of BN in an application merely amounts to using a set of images to derive simple statistics for the conditional probabilities. Consequently, compared to discriminant-based systems such as neural networks, which are vulnerable to poor generalization because they tend to memorize the training set, testing the BN-based systems generally does not give overly biased results.

E. Remarks

The benchmark research in this paper shows that the neural networks perform worse than BN. In other words, the BN-based systems have more excellent generalization than NN because NN tends to memorize the training set to a certain degree. At that time researchers usually think that increasing the size of the training set and imposing additional constraints on the neural network training method (such as a validation stop) can mitigate the memorization effect but will result in reduced performance from the neural network on the training set of images.

However, the year after this paper was published, the deep neural networks were proposed and perform better than the solely BN framework on the field of semantic image understanding. Meanwhile, a combination of different classifiers in an ensemble are becoming popular in order to improve accuracy.

A generative method has been proposed to combine implicit and exploit knowledge in the second paper using the Bayesian Network. Moreover, the focus-of-attention mechanism has improved the efficiency on verifying the hypotheses by the BN and significantly reduce the computational cost of visual inference while obtaining results comparable to the exhaustive case.

III. Evidence-Driven Image Interpretation by Combining Implicit and Explicit Knowledge in a Bayesian Network

In contrast to human perception that makes extensive use of logic-based rules, the models, normally relying on a set of discriminative features, fail to benefit from knowledge that is explicitly provided. As for the multimedia, the use of domain knowledge has been motivated o index this type of data by the difficulty of mapping a set of low-level visual features into semantic concepts. Furthermore, owing to the importance of context in understanding audiovisual stimuli, the integration of context and content is considered a promising approach toward multimedia understanding [4].

In this paper [5], the authors firstly combined ontologies and BNs to allow, in a probabilistic way, the fusion of evidence obtained at different levels of image analysis. They proposed a data-oriented learning strategy to estimate the parameters of the BN. Secondly, they showed how global and regional evidence which were obtained from the application of concept classifiers on global and local image data can probabilistically be combined within a BN. Combining information this way is demonstrated to lead to statistically significant improvements for the three tasks: image categorization, localized region labeling and weak annotation of video shot keyframes.

A. Framework Introduction

The framework description consists of five parts: Visual stimulus, Domain knowledge, Application context, Evidence-driven probabilistic inference and the Computational efficiency.

As for the visual stimulus, the authors consider the supervised learning paradigm to analyze it. In the learning paradigm, a classifier is trained to identify an object category and provided that a sufficiently large number of examples are available. When Fc is a probabilistic classifier, we have $F_c(I_q) = Pr(c|I_q)$, where I_q is the analyzed visual representation and c is an instance of the concept.

The authors used use OWL-DL to build the structure of the domain knowledge: $K_D = S(N_C, R, O)$. It describes how the domain concepts are related to each other using R and $O \in DL$, where DL stands for "description logics" [6] and constitutes a specific set of constructors. The use of the structure is to trigger the probabilistic inference process and to know which evidence supports a certain hypothesis and what semantic restrictions apply in this domain. The knowledge structure sets the tracks to which evidence belief [7] is allowed to propagate by determining the structure of the BN. The information, which is implicitly extracted from the training data, is encoded into the CPTs of the BN nodes and influences the probabilistic inference process when belief propagation takes place.

The authors considered the application context as $X = S(app, W)$, it is the information that consists of both app, the type of application-specific information, and $W = [W_{i,j}]$, whose elements $W_{i,j}$ quantifies the effect of concept c_i on c_j. It specifies the quantitative relations between evidence and hypotheses, expressed with W.

To accommodate for evidence-driven probabilistic inference, the framework uses a BN derived from the domain ontology. Let $h(I_q, c_i) = Pr(c_i|I_q)$, it is the function to estimate the degree of confidence that concept c_i appears in image I_q. Moreover, the set of confidence degrees that the concepts that belong to the hypotheses set are depicted in image I_q can be expressed as $H(I_q) = h(I_q, c_i) : c_i \in c^H$. And $E(I_q) = h(I_q, c_i) : c_i \in c^E$ represent the set of confidence degrees that the concepts that belong to the evidence set are depicted in image I_q. In the model, $H(I_q)$ and $E(I_q)$ are provided to the BN. Using the probabilistic reference, the posterior probabilities of the network nodes are calculated by the information of the knowledge of R, O, and context $W_{i,j}$. The authors demonstrated the framework by $h^{'}(I_q, c_i) = Pr(c_i|H(I_q), E(I_q), R, O, W_{i,j})$ which is used to calculate the posterior probabilities of the network nodes. Furthermore, the set of posterior probabilities of the concepts that belong to the hypotheses set can be represented as $H^{'}(I_q) = h^{'}(I_q, c_i) : c_i \in^H$. And the semantic image interpretation is expressed as $c = arg \oplus h^{'}(I_q, c_i)$.

The computational cost for gathering the necessary evidence is often very expensive, which can be prohibitive in highly complex domains. Therefore, the authors used an original method called a FoA mechanism to improve the computational efficiency of the proposed framework. This

method is performed by calculating the mutual information between the node that corresponds to the concept c_k and all other nodes that correspond to the concepts of c^E, where c_i is a ranked list of the evidence concepts (i.e., $\forall c_i \in c^E$). The mutual information between c_k and $c_i, \forall c_i \in c^E$, is calculated according to the following equation:

$$I(c_k; c_i) = \sum_{\{true,false\}} \sum_{\{true,false\}} Pr(c_k; c_i) log_2 \frac{Pr(c_k; c_i)}{Pr(c_k)Pr(c_i)}, \quad (8)$$

where $Pr(c_k, c_i)$ is the joint, and $Pr(c_k), Pr(c_i)$ are the marginal probability distributions of c_k and c_i, respectively. The efficient calculation of $Pr(c_k, c_i)$ is performed using the junction tree [8], which is an efficient and scalable belief propagation algorithm that exploits a range of local representations for the network.

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 $H^{'}(I_q)$ 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 can encode the qualitative characteristics of causality (i.e., which nodes affect which) and the 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 can facilitate the following three different operations:

- Providing the means to store and utilizing domain knowledge K_D, which is achieved by mapping K_D to the network structure;
- Organizing and make accessible information from the application context $W_{i,j}$, which is achieved by the CPTs attached to the network nodes;
- Allowing the propagation of evidence belief in a mathematically coherent manner, which is performed with the use of message-passing belief propagation algorithms.

B. Main Method

1 Defining K_D, app, c^E as introduced before.
2 Applying the probabilistic classifiers F_c on I_q to obtain the degrees of confidence for the concepts in c^E .
3 Using app and K_D to decide which of the domain concepts should constitute the hypotheses set c^H .
4 Providing the degrees of confidence for the concepts in c^E to the BN and trigger probabilistic inference by using these degrees as soft evidence.
5 Propagating evidence beliefs using the network's inference tracks R and the corresponding causality quantification functions $W_{i,j}$. The conditional probabilities are learned by employing the expectation-maximization (EM) [9] algorithm, using as training data the images annotated with concept labels.
6 Using a FoA mechanism. It is based on the mutual information between concepts which 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.
7 Calculating the posterior probabilities for all concepts in c^H and decide which of the hypotheses should be verified or rejected. To find the concept that matches best a given image is to use a greedy search method by alternating on the set of hypotheses and the set of evidences in order.

C. Ontologies and Bayesian Networks

In this paper, the authors used data from two datasets: the "Personal Collection" (PS) and the "News". The 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. As can be seen from the ontology, used to represent the PS domain knowledge and deriving the BN, is shown in Fig. 6, the set of category concepts is $C_G = \{Countryside, buildings, seaside, rockyside, forest, tennis, roadside\}$ and the set of regional concepts is $C_L = \{Building, roof, tree, stone, grass, ground, dried-plant, \cdots\}$.

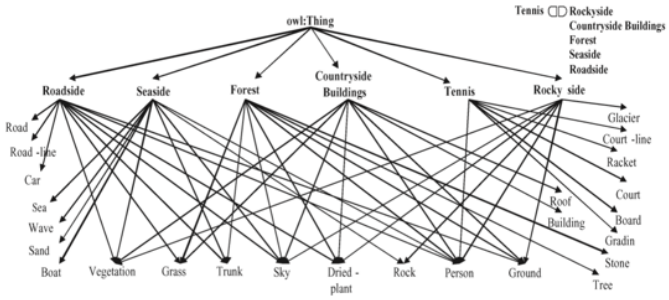

Figure 6: Ontology for the "Personal Collection" domain

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, i.e., $H(I_q) = \{Pr(c_i|I_q) : i = 1, \cdots, n\}$, where n is the number of category concepts in K_D. For example, knowing that a specific region depicts a road is a type of contextual information that the algorithm can exploit when trying to decide whether the image depicts a rocky side or a roadside scene. To assess the benefit of using the proposed FoA mechanism, we measure the gain in computational cost in terms of the following two quantities: 1) the number of classifiers (#Classifiers) that need to be applied and 2) the number of inferences (#Inferences) that need to be performed. The authors set

three baselines: $CON1$, $CON2$ and $CON3$. In the baseline configuration $CON1$, they assessed 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., $K_D = S(N_C, R, O)$) to extract the existing evidence and facilitate the process of evidence-driven probabilistic inference. The third configuration $CON3$ takes into account the semantic constraints of the domain to construct the BN. The Fig. 7 shows that the performance obtained using the $CON2$ is superior to the performance obtained using $CON1$, because an average increase of approximately 5% is observed.

Localized region labeling is the task of assigning labels

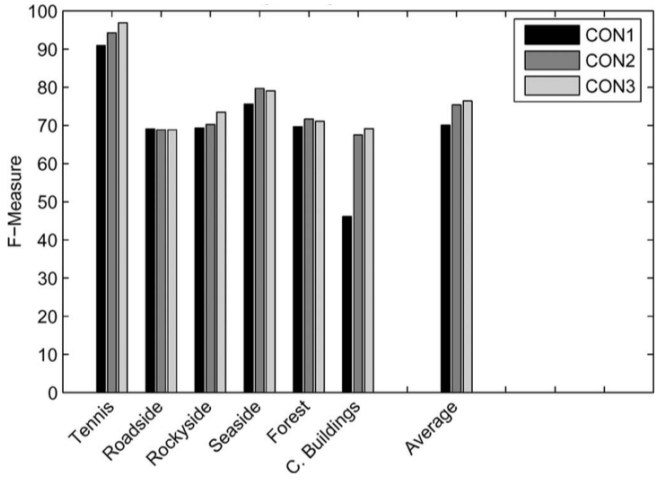

Figure 7: Image Categoriaztion Evaluation

to presegmented image regions with one of the available regional concepts c_i'. Eventually, the values that represent the impact on the posterior probabilities of the two different cases are compared. If no conflict occurs, the concept that corresponds to the local classifier with maximum confidence is selected. Fig. 8 shows that, when using the proposed framework, an average increase of approximately 4.5% is accomplished.

Weak annotation of video shot keyframes is the task of associating a number of concepts with an image. The evidence are considered the confidence values of all other concepts, except for the concept examined by the current hypothesis. To assess the efficiency of the framework, the authors compared its performance to the performance of baseline concept detectors that make no use of domain knowledge and application context. Belief propagation is performed, and the resulting posteriors are recorded for all concepts.

The experiments conducted have verified the effectiveness of our framework in improving the performance of a set of baseline concept classifiers by using their output as evidence. Because this improvement mainly derives

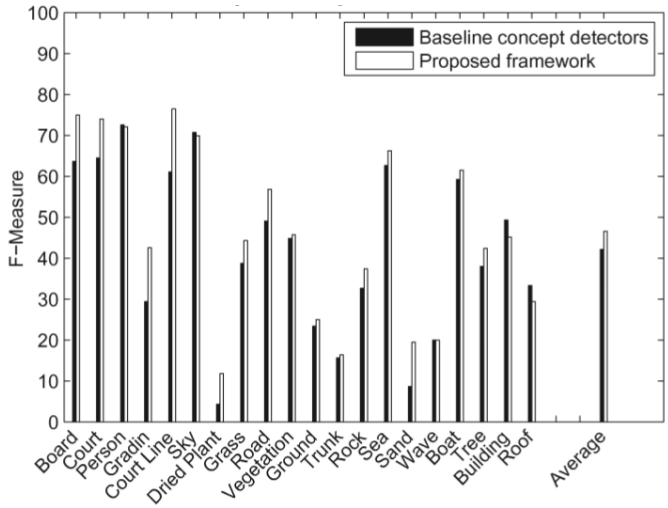

Figure 8: F-Measure scores for the localized region labeling task-Personal Collection Dataset

from the incorporation of the domain knowledge and the application context to the analysis, we can use the proposed framework to improve the performance of any set of concept detectors that produce a probabilistic output. We see that the FoA mechanism is as successful as a full brute force classification search. Using ontologies thus reduces the amount of computation while giving good results.

D. Remarks

The experiments have conducted and verified the effectiveness of the framework in improving the performance of a set of baseline concept classifiers by using their output as evidence. Derived from the incorporation of the domain knowledge and the application context to the analysis, the framework can be used to improve the performance of any set of concept detectors that produce a probabilistic output. Furthermore, the mechanism that exploits the mutual information between concepts to significantly reduce the computational cost of visual inference and still achieve results comparable to the exhaustive case.

IV. Conclusion and Discussion

In the first paper, the general framework for semantic image understanding presented above can be applied to various tasks involving semantic understanding of pictorial images. With these diverse examples, they have demonstrated that effective inference engines can be built according to specific domain knowledge and available training data to solve inherently uncertain vision problems. BN are becoming the reasoning engine of choice and provide a powerful tool applicable to many photograph-related semantic understanding tasks.

The second paper was written in 2011 and proposed a generative method of modeling the layer of evidence to

effectively combine and exploit both a priori and observed information. The authors combined everything in a Bayesian network that can perform inference based on soft evidence and provided the means to handle aspects such as causality, uncertainty, and prior knowledge, hence imitating some human basic perceptual operations when inspecting images.

These two papers have similar topic but different methods. Both demonstrate how to make good use of various features or evidences contained in images to infer high-level semantic interpretation under BN framework. And their common 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.

The BN frameworks as well as training methods they developed are distinct. For BN frameworks, in the first paper, two sets of descriptors are extracted from the image: the first set corresponds to low-level features, such as color, texture, and edges; the second set corresponds to semantic objects that can be automatically detected. Then, the hybrid streams of low-level and semantic evidences are piped into a BN-based inference engine, which is capable of incorporating domain knowledge as well as dealing with a variable number of input evidences, producing semantic predicates. In the second paper, an ontology was used to represent the domain knowledge. Global classifiers are applied to estimate the initial probability for each hypothesis. and local classifiers are applied to the presegmented image regions I_q to generate a set of confidence values that constitute the evidence. For training methods, in the first paper, the CPMs (Conditional Probability Matrices) are computed using fractional frequency counting, while in the second paper the conditional probabilities are learned by employing the EM algorithm, and the CPTs of all control nodes are manually set.

REFERENCES

[1] L. N. Matos and J. M. D. Carvalho, "Combining global and local classifiers with bayesian network," in *International Conference on Pattern Recognition*, 2006, pp. 1212–1215.

[2] Z. Ding, Y. Peng, and R. Pan, "A bayesian approach to uncertainty modelling in owl ontology," *A Bayesian Approach to Uncertainty Modelling in Owl Ontology*, 2006.

[3] J. Luo, A. Singhal, S. P. Etz, and R. T. Gray, "A computational approach to determination of main subject regions in photographic images," *Image & Vision Computing*, vol. 22, no. 3, pp. 227–241, 2004.

[4] A. F. Smeaton, "Content vs. context for multimedia semantics: The case of sensecam image structuring," in *International Conference on Semantic and Digital Media Technologies*, 2006, pp. 1–10.

[5] S. Nikolopoulos, G. T. Papadopoulos, I. Kompatsiaris, and I. Patras, "Evidence-driven image interpretation by combining implicit and explicit knowledge in a bayesian network," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 41, no. 5, pp. 1366–1381, 2011.

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

[7] 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.

[8] F. V. Jensen and F. Jensen, "Optimal junction trees," *Uai*, pp. 360–366, 2013.

[9] G. J. Mclachlan and T. Krishnan, "The em algorithm and extensions," *Biometrics*, vol. 382, no. 1, pp. 154–156, 1997.