PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT

Pattern Recognition 38 (2005) 919–934

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

A Bayesian network-based framework for semantic image understanding

Jiebo Luo[a,*], Andreas E. Savakis[b], Amit Singhal[a]

[a]*Research and Development Laboratories, Eastman Kodak Company, 1850 Dewey Ave., Rochester, NY 14650-1816, USA*
[b]*Department of Computer Engineering, Rochester Institute of Technology, 83 Lomb Memorial Dr., Rochester, NY 14623, USA*

Received 9 August 2004

Abstract

Current research in content-based semantic image understanding is largely confined to exemplar-based approaches built on low-level feature extraction and classification. The ability to extract both low-level and semantic features and perform knowledge integration of different types of features is expected to raise semantic image understanding to a new level. Belief networks, or Bayesian networks (BN), have proven to be an effective knowledge representation and inference engine in artificial intelligence and expert systems research. Their effectiveness is due to the ability to explicitly integrate domain knowledge in the network structure and to reduce a joint probability distribution to conditional independence relationships. In this paper, we present a general-purpose knowledge integration framework that employs BN in integrating both low-level and semantic features. The efficacy of this framework is demonstrated via three applications involving semantic understanding of pictorial images. The first application aims at detecting main photographic subjects in an image, the second aims at selecting the most appealing image in an event, and the third aims at classifying images into indoor or outdoor scenes. With these diverse examples, we demonstrate that effective inference engines can be built within this powerful and flexible framework according to specific domain knowledge and available training data to solve inherently uncertain vision problems.
© 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Semantic image understanding; Low-level features; Semantic features; Bayesian networks; Domain knowledge

1. Introduction

Low-level features, such as color, texture, and shape, have been widely used in content-based image processing and analysis [1–3]. While low-level features are effective for certain tasks, such as "query by example", they are rather limited for many multimedia applications, such as efficient browsing and organization of large collections of digital photos and videos, that require advanced content extraction and image understanding [3]. Therefore, the ability to extract semantic features in addition to low-level features and to perform fusion of such varied types of features would be very beneficial for scene interpretation.

Since a large number of semantic understanding tasks are performed on photographic images, it is important to understand some of the fundamental characteristics of photographs: (a) they have unconstrained picture content and are taken under unconstrained imaging conditions; (b) they serve the purpose of recording and communicating memories and, therefore, enable a certain degree of consensus among first- and third-party observers with respect to the intent of the photographer; (c) the feature extraction process is imperfect, due to limitations in the accuracy of feature extraction algorithms, as well as the limitations in our understanding of the problem.

* Corresponding author. Tel.: +1 585 722 7139; fax: +1 585 722 0160.
E-mail address: jiebo.luo@kodak.com (J. Luo).

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2004.11.001

Due to the unconstrained nature of photographic images, and the lack of fully reliable low-level features, it is advantageous to select a diverse set of features that extend beyond the standard color/texture/shape types. Semantic features are excellent candidates for providing diversity in the feature set and their use has been proposed in addition to low-level features. The challenge with such an approach is that knowledge from diverse feature sets needs to be integrated, so that specific inferences can be made. In addition, the inference engine should be capable of resolving conflicting indicators from various features, which are likely to occur due to the imperfect nature of the feature extraction algorithms. A unified framework for integrating both low-level and semantic features would be extremely valuable for image understanding, because it would allow for diversity in the feature extraction process and the incorporation of features that are different in nature.

Classic work on feature fusion can be categorized primarily into rule-based methods, voting methods, and discriminant-based methods. Rule-based methods require that the rule designer have knowledge of all possible conditions, which allows for the design of complete and efficient rules [47,4,5]. Rule-based methods can be effective in restricted environments, however, the unconstrained nature of photographs makes it difficult to effectively employ rule-based methods in general situations. Fuzzy logic-based algorithms can also be considered in this category. Voting methods can be as simple as majority voting, or they may involve more sophisticated weighting approaches, which in some cases resemble rule-based methods [6–8]. The difficulty with voting methods lies in determining the weights of different feature types. While it is convenient to assume equal weights for all features, in practice it becomes necessary to adjust the weights according to feature type, as dictated by the application on hand. Discriminant-based methods include neural networks and other equivalent classifiers, e.g., fuzzy neural networks [9–11] and support vector machines (SVM) [12,13], which treat all features as combined vectors. The difficulty with the feature vector approach is the lack of insight into how each feature influences the combined decision. Nevertheless, discriminant-based methods have received considerable attention and have proved effective in a variety of applications [12,13,16,17].

In this paper, we present a framework for semantic image understanding based on belief networks. The framework is suitable for applications where semantic understanding of pictorial images is important. Three examples are presented as case studies for using the proposed framework. The first example is main subject detection (MSD), i.e., determining the likelihood of a given *region* in an *image* being the main subject. Another example is emphasis image selection (EIS), i.e., selecting the most appealing *image* in an *event* comprising of a number of related images. The final example is scene classification of images into indoor or outdoor scenes. The proposed general framework and Belief networks are discussed in Sections 2 and 3. The applications are outlined in Section 4, followed by a benchmarking study of Bayesian networks (BN) vs. neural networks for MSD in Section 5 and conclusions in Section 6.

2. A general framework for image understanding

2.1. Review of existing image understanding frameworks

Image understanding is the process of converting "pixels to predicates", i.e., iconic image representations to symbolic form of knowledge [21]. Image understanding is the highest (most abstract) processing level in computer vision [22], as opposed to image processing, which converts one image representation to another, for instance, converting raw pixels to an edge map.

Much of the early successes in image understanding have been made in constrained environments, e.g., automatic military target recognition [23], and document [24] and medical [25] image understanding. While image understanding in unconstrained environments is still largely an open problem [16,22], progress is being made in scene classification where the goal is to place an image into one of a set of *predefined* physical (e.g., indoor or outdoor, upright or upside-down image orientation) or semantic categories (e.g., beach, sunset).

Complete object recognition is not necessary and often not possible, especially given the current capabilities of computer vision systems. Fortunately, scenes can be classified without full knowledge of every object in the image. It may be possible, in some cases, to use low-level information, such as spatial distribution of color and texture, to classify some scene types with a high level of accuracy.

One major approach to semantic image understanding is based on the above premise. Examples or training data are collected. These exemplars are thought to fall into clusters in the feature space. They are used to train an appropriate classifier to classify novel test images. In essence, exemplar-based approaches apply pattern recognition techniques (discriminants) to vectors of low-level image features (such as color, texture, or edges) and semantic image understanding is achieved according to the similarity between novel test images and the training exemplars according to a distance metric measured in the selected feature space.

Exemplar-based systems using low-level features have demonstrated successes as well as limitations. Low-level features have the advantage of simplicity. Global or local features are calculated for each image without having to first segment the image or recognize objects in the scene, which can be as challenging as image understanding itself. For tasks such as indoor–outdoor image classification [1], scene classification [2], and image orientation detection [12,17], respectable performance has been achieved. However, even for the same tasks, higher level features or cues are clearly demanded. For instance, some natural images pose difficulty even for a human to decide the correct orientation at a low

resolution where object recognition is difficult or impossible, and some may not even have a preferred orientation. Another major concern with exemplar-based approaches is the generalizability to real-world, unconstrained images, which do not fall into well-defined scene prototypes, and for which a comprehensive collection of prototype exemplars is not possible.

Model-based approaches are built on the expected configuration of a specific type of scene. A scene's configuration is the layout (relative location and size) of its objects, created from expert knowledge of the scene. Relatively little research has been done on using model-based approaches for unconstrained natural image understanding, because it is usually only possible to build a model for a well-defined scene type, and such a model is usually not generalizable to other scene types [26,27]. For example, while it is possible to build scene models *manually* and *individually* for scene types such as "fields", "snowy mountains", "snowy mountains with lakes", and "waterfalls", it would be far more difficult to do so for other scenes types such as typical indoor scenes. More recently, a trainable scene configuration model called composite region template was proposed in Ref. [28] and shown to be promising for a selected set of scene types exhibiting distinctive spatial configuration patterns.

In many cases, *selective* object recognition is possible and helpful, even though it is impossible to design detectors for *every* object in the scene [18,19]. This is the main motivation for the proposed general framework in which detectable semantic features, which are effective but often difficult to obtain, and low-level features, which are always available but not as effective, are combined to provide a *rich* description of the scene. We argue that by using low-level vision features to *complement* a small number of selected semantic object detectors it is not necessary to detect every object in a scene to achieve a descent performance for a specific task.

The proposed framework may be viewed as a hybrid approach. First, both low-level and semantic features are utilized. In fact, it is a great challenge to find a way to combine such diverse information, measured by different metrics, and represented by different means. For example, color features are represented by histograms, and the presence of a face is Boolean. A probabilistic knowledge integration framework allows all the information to be integrated in common terms of probabilities. Second, both bottom-up and top-down control strategies may be used; low-level features are usually used in a bottom-up fashion by classifiers, while semantic features may have to be derived using top-down models. Third, domain knowledge is crucial for the visual inference process, because it can help bridge the "sensory gap" and "semantic gap" [3], and serve as the catalyst for fusing low-level and semantic features. We argue that *domain knowledge* is extremely important in making up for the deficiencies in bottom-up approaches and more practical to use than top-down approaches for unconstrained natural images. BN allow domain knowledge to be incorporated in the structure as well as the parameters of the networks, which is more difficult, if not impossible, for other inference engines such as neural networks or SVM.

2.2. The 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). The state of the art in object detection, both in terms of accuracy and speed, puts a limit on what is included in the object detector bank. 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, which may be in the form of semantic labels of the entire image or importance maps indicating different scene content.

In summary, the proposed framework is advantageous because it is capable of accommodating the following factors:

- the successes and limitations of low-level feature-based approaches;
- the need to integrate selected semantic features when available;
- the need to integrate critical domain knowledge;
- the need to make use of limited ground truth training data (generalizability).

Bayesian belief networks form the backbone of the feature integration framework by providing explicit probabilistic capabilities for representing diverse feature sets in a common modality (probability space) and theoretically sound fusion rules for combining these probabilities to generate a consensus. Before we describe the use of this framework in selected photographic image understanding applications, it is important to understand the basis of Bayesian belief networks and their use for data fusion and classification.

3. Bayesian belief networks

3.1. Bayesian networks (BN)

Belief networks, or Bayesian nets, have proved to be an effective knowledge representation and inference engine in artificial intelligence and expert systems. Domain-specific

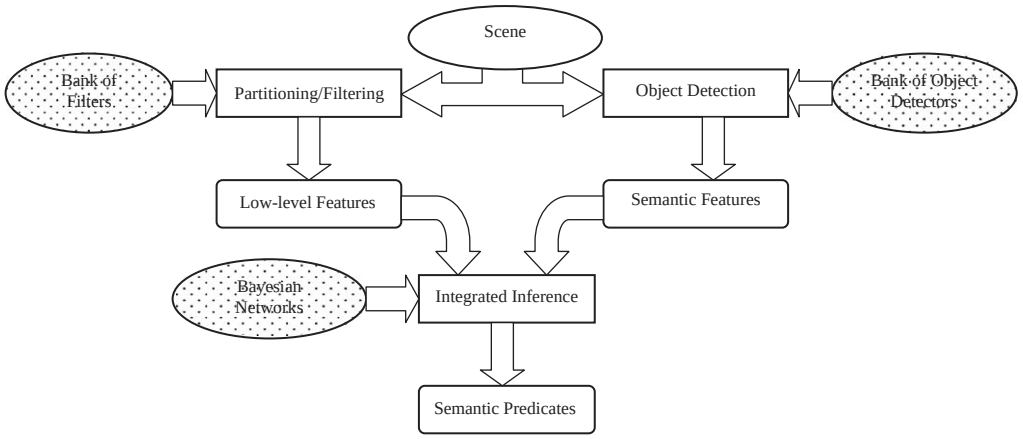

Fig. 1. A general framework for semantic image understanding.

knowledge can be incorporated in the network structure, and a complicated joint probability distribution can be reduced to a set of conditionally independent relationships that are easier to characterize. Thus, a BN can be used to represent the dependence between random variables (features) and to give a concise and tractable specification of the joint probability distribution for a domain.

BN are directed, acyclic graphs that encode the cause–effect and conditional independence relationships among variables in the probabilistic reasoning system [15]. The directions of links represent causality. The links between the nodes, or variables, represent the conditional probabilities of inferring the existence of one variable (the destination of the link) given the existence of the other variable (the source of the link). Each node can have many such directed inputs and outputs, each specifying its dependence relationship to the nodes from which the inputs originate (parents) and nodes where the outputs go (children).

According to Bayes' rule, the posterior probability can be expressed by the joint probability, which can be further expressed by the conditional and prior probabilities:

$$P(S \mid E) = \frac{P(S, E)}{P(E)} = \frac{P(E \mid S)P(S)}{P(E)}, \quad (1)$$

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. However, as the number of variables grows, the joint probability can become intractable. With BN networks, the computation of the joint probability distribution over the entire system given partial evidences about the state of the system is greatly simplified by using Bayes' rule to exploit the conditional independence relationships among variables.

A Bayes net 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.

A Bayes network can be viewed as a knowledge representation mechanism because it encodes the joint probability distribution. It can also be considered as an inference engine because its evaluation produces the posterior joint probability distribution given evidence of various variables. Its advantages include explicit uncertainty characterization, fast and efficient computation, quick training, high adaptivity and ease of use, and explicit representation of domain-specific knowledge akin to a human reasoning framework. We found that for our applications, Bayes networks offer good generalization with limited training data, easy maintenance when adding new features or new training data, and convenience in building performance-scalable versions by pruning features. Other researchers also confirmed that such a frequency counting-based inference engine provides good generalization because the outcome of BN is usually insensitive to small disturbances in the frequency counts [29,30]. Since BN derive much of their power from their ability to reduce a complex joint probability distribution into a series of simpler conditional probability distributions, it is worthwhile to spend some effort understanding the capabilities and limitations imposed by the conditional independence assumption.

3.2. On conditional independence

There has been a growing interest in recent years in developing real world applications of BN due to their knowledge representation and inference capabilities. This has been accompanied by development of effective techniques for knowledge engineering as well as efficient algorithms for performing inference using BN. The belief propagation al-

gorithm for a general BN is NP-hard. To make the BN framework tractable for use in practical applications, various approximations have to be made depending on the application. This could include assumptions about the structure of the network, the independence relationships of the network, or imprecise conditional and prior probabilities.

Pradhan et al. [29] and deDombal et al. [30] provide an excellent review of various empirical studies that have been performed to determine the sensitivity of BN to imprecise probabilities and approximations of independence assumptions. Kwoh and Gillies [31] show that the effect of ignoring dependencies in a BN model is dependent on a number of factors associated with the variables. Suppose two variables measured the same property, and were included in the BN without resolution of this dependency. Also, suppose that these were very effective variables in that they provided for strong inferences of the parent variable given positive evidence. In this case, the positive evidence could be magnified to such a large degree that it could lead to incorrect inferences about the parent variable. However, if the two variables did not have a strong causal relationship with the parent variable, then their effect on the posteriors of the parent variable would not be so drastic.

In practical applications of BN, it is necessary to employ certain restrictions and assumptions to make the problem of belief propagation in the BN tractable [29,31,32]. Fung and Del Favero [32] advocate using imprecise independence assumptions to simplify the BN structure for the problem domain of information retrieval. They use a single-level Bayesian network (SLBN) to relate the occurrence of certain keywords in an article (e.g., the words stock, portfolio, profit, etc.) to the topic of that article (e.g., investment). The BN then has, as child nodes, all the various words identified as features for a given topic, which is the root node. The occurrence of various words (features) in an article of given topic is considered to be independent of each other. The authors suggest that this is a reasonable assumption to make if synonyms or antonyms are not used as multiple features. However, from the example topic given earlier in this discussion, the features “stock”, “portfolio”, and “profit” are not synonyms or antonyms, but are definitely not independent of each other. The reported results show that this type of imprecise conditional independence can still lead to a practical system with good accuracies.

Sarkar and Chavali [33] use a similar set of constraints for constructing a BN for modeling parameter space behavior of vision systems. They look at the problem of assigning an optimal set of parameters and thresholds to various vision tasks. The BN capture the conditional dependencies between various parameters in the system and are learned from data. They restrict the BN to a polytree structure that, obviously, leads to a less than accurate representation of the conditional dependencies of the variables in the network. They claim the imposition of a polytree structure is convenient from a computational standpoint and provides reasonable results. The results show that the BN generated parameter sets result in good performance for the vision systems and outperform parameter sets generated using some simple heuristics. Also, they claim that 10 images provided sufficient training data for the BN resulting in similar performance on the train and test image sets. Kwoh and Gilies [31] propose the use of *Pearson's correlation coefficient* to test the conditional dependencies among various variables in a system. For example, if there are three variables $\{A, B, C\}$ and we want to test whether B and C are conditionally independent given A, i.e.

$$P(A, B, C) = P(A)P(B|A)P(C|A) \tag{2}$$

we can compute the Pearson's correlation coefficient, ρ using the following equation:

$$\rho = \sum P(A)\frac{Cov(B|A, C|A)}{\sqrt{var(B|A)var(C|A)}}. \tag{3}$$

In our framework, we have chosen to adopt a mixture of the above approaches. We carefully analyze the variables in our system when building the BN. Variables whose conditional independence cannot be ascertained using semantics are further subjected to a rigorous testing of the dependencies using Pearson's correlation coefficient. One note of caution here is that blindly using Pearson's correlation coefficient may lead to the encoding of coincidental data dependencies (data correlation) that may be an artifact of the training data and not truly reflect the causal dependencies in the system. Once the BN structure has been determined (by a diligent analysis of the independence relations between variables in the domain), we can use automatic methods to train the parameters associated with the network structure. It is important to note that while we advocate the use of a domain expert to define and model the independence relationships for a given application, there exist approaches for automatically learning the structure of a BN from training data [34,35]. However, these methods tend to construct the network based on correlation rather than causation.

3.3. Training the Bayesian network parameters

Various researchers have proposed schemes for learning the parameters (CPMs, priors, etc.) associated with a BN [34,35]. The most common of these is to use Bayesian statistics to learn the parameters of a given network structure from data. When the observed data is complete, this reduces to a simple frequency counting approach where likelihoods of observing a set of variables are generated from the training samples (the observed data).

There are two standard methods for obtaining the CPMs for each parent-child node pair (or tree in case of multiple parents): (1) expert knowledge, and (2) frequency counting. Expert knowledge-based training is a knowledge-engineering method where an expert is consulted about the

relationship between the label sets of the two nodes joined by each link. Using this knowledge, the CPM for each node pair can be generated. The expert can be the designer of the network, who may have intimate knowledge of the relationships between various entities in the domain, or a recognized scholar in the problem domain. It is apparent that the CPMs obtained in such a manner are only as reliable as the expert. However, if the desired conditional relationships are well understood or the training data is either extremely hard or even impossible to obtain, then expert knowledge-based training of the network may be the best option. Frequency counting-based training is a sampling and correlation method that can be used for learning the CPMs directly from training data. In this case, a large set of observations (of the state of each variable in the network) and ground truth is first collected. The conditional probability matrix for a link can be trained using frequency counting only when ground truth for the parent node is available. Ground truth, in its normal sense, refers to knowing the label (associated with the parent node of the link) of each training sample with absolute certainty. Multiple observations of each child node are recorded along with ground truth on the parent node. These observations are then compiled together to create frequency tables which, when normalized, can be used as the CPM. The frequency, f, is computed as follows:

$$f = \frac{\sum_{r\in R} T_r e_r}{\sum_{r\in R}}, \tag{4}$$

where R is the set of all regions, T_r represents the ground truth value for region r, and e_r represents the feature detector output for region r. The above equation works when the ground-truth is binary in nature. We use a slightly modified notion of ground truth where the parent-node label of each training sample is known with some degree of (instead of absolute) certainty. This is the case in applications such as MSD and EIS where a number of independent observers were used to collect individual ground truth on the same set of images for an inherently subjective task. The final ground truth is compiled by averaging the individual ground truth to create a composite probabilistic ground truth.

To address this problem of partially certain ground truth, we have developed a *fractional frequency counting*-based training method. Fractional frequency counting-based training is very similar to the frequency-counting approach except that we now weight the feature measurements using 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, we 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:

$$CPM = \left[\left(\sum_{i\in I}\sum_{r\in R_i} n_i F_r^{\mathrm{T}} T_r\right) C\right]^{\mathrm{T}}, \tag{5}$$

$$F_r = [f_o^r f_1^r f_2^r \ldots f_m^r], \quad T_r = [t_o^r t_1^r t_2^r \ldots t_l^r], \tag{6}$$

$$C = diag\{p_j\}, \quad p_j = \left(\sum_{i\in I}\sum_{r\in R_i} n_i t_r\right), \tag{7}$$

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.

It is easy to show that our probabilistic ground truth formulation, where the ground truth represents the average belief of the observers in each region being a main subject region, is equivalent to the traditional approach where each observation (from each observer) is treated individually. In the latter case, 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

$$f = \frac{\sum_{o\in O}\sum_{r\in R} T_{o,r} e_r}{\sum_{o\in O}\sum_{r\in R}}, \tag{8}$$

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 ($N = \sum_{o\in O}$), the above equation is equivalent to

$$f = \frac{\sum_{r\in R}\sum_{o\in O} T_{o,r} e_r}{N\sum_{r\in R}} \tag{9}$$

which is equivalent to

$$f = \frac{\sum_{r\in R}(\sum_{o\in O} T_{o,r}/N) e_r}{\sum_{r\in R}}. \tag{10}$$

The term $\sum_{o\in O} T_{o,r}/N$ is the probabilistic ground truth.

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 three different applications in the photographic image understanding domain. These applications are MSD, EIS, and indoor–outdoor classification.

4. Applications

4.1. Main subject detection (MSD)

In photographs, the photographer intends to convey to the viewer his or her interest in one or more main subjects.

Mechanisms for doing so include controlling the position of the main subject within the frame, arranging for a contrast in tone, color, or texture between the main subject and its surroundings, and highlighting the primary subject. In general, third-party viewers are in agreement as to what is the main subject in a particular picture, and to a large extent their judgment reflects the intention of the photographer [36]. The goal of MSD is to develop automatic algorithms that identify, from the perspective of a third-party viewer, the main subject in a photograph. This is done in a statistical sense that reflects the degrees of ambiguity inherent to such a task [36].

We view MSD as a measure of saliency or relative importance for different image regions that, after segmentation, may be associated with different subjects in an image. It enables a discriminative treatment of the scene content for applications in image understanding, enhancement, and manipulation. It is also related to the topic of automatic detection of interesting or important regions in an image [37].

We have developed a probabilistic reasoning approach to MSD [20]. 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 [20], 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 $P(Background)=0.72$.

The BN was constructed based on the perceived semantic relationships between features used for MSD. In addition, Pearson's correlation coefficient was computed for feature pairs identified as conditionally independent as a secondary check on the structure. A number of currently available feature detectors were analyzed for interdependencies. We first built a simple SLBN using only those features identified as conditionally independent of each other. This network served as a baseline performer. Since a number of features were discovered to have dependencies that could not be accounted for in a SLBN, we also constructed a multi-level Bayesian network (MLBN) that introduced *hidden* variables to make these features conditionally independent. In the MLBN, features that are not conditionally independent given *MainSubject* are organized in separate subtrees. The root nodes of the subtrees (e.g., *GeometricShape*, *KeySubjectMatter*, and *Location*) are called intermediate nodes because they link the feature detectors to the *MainSubject* node in the multi-level Bayes net. Other major intermediate nodes are *Convexity* and *RelativeShape*. For example, the *GeometricShape* subtree combines evidence from *Rectangularity* and *Circularity*, two shape feature detectors that are not conditionally independent when *MainSubject* is known.

The MLBN is constructed by linking these subtrees to the *MainSubject* (root) node of the network. Feature detectors that are completely independent (e.g., *Size*, *Symmetry*) are directly linked to the root node of the network. This results in a tree-like structure for the Bayes network shown in Fig. 2.

A second issue arises from the use of multiple detectors for the same feature. As an example, we have three feature detectors measuring the convexity of the region. Including all three in the SLBN leads to triple counting of convexity evidence. Therefore, only one of the convexity feature detectors can be used in the single-level network. In the MLBN, an intermediate node, *Convexity*, is used to gather evidence about a particular feature from multiple detectors and combined to form single evidence that is then used for inference at the top-level.

Experimental results were obtained for 100 images carefully selected to match the characteristics of a typical "photospace" in terms of the expected frequencies of occurrence of indoor/outdoor, people, landscape, subject distance, lighting, etc. Half of the images were used for training and the other half for testing. The associated belief generated by the Bayes net is attached to each region, so that regions with large values correspond to regions with high confidence or belief in being part of the main subject. This reflects the inherent uncertainty for humans to perform such a task.

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. A detailed benchmarking of various classifiers and predictors for MSD was also performed. The results of the benchmarking study and the performance of the BN-based system for MSD are described in further detail in Section 5.

4.2. Emphasis image selection (EIS)

In a variety of applications that deal with a group of pictures, it is important to rank the images in terms of their relative appeal, called emphasis, so that they can be processed or treated according to their emphasis value. EIS is defined

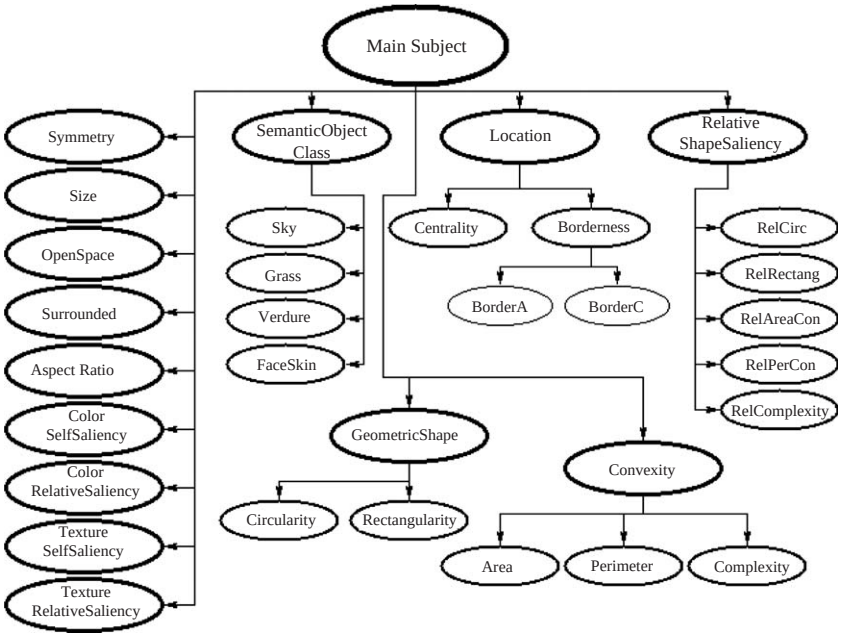

Fig. 2. A BN for main subject detection.

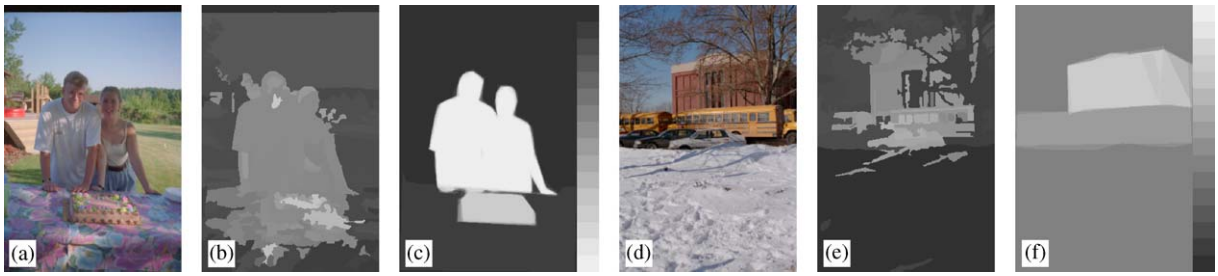

Fig. 3. Examples of MSD results: (a,d) images; (b,e) MSD results; (c,f) ground truth maps.

as follows: given a set of photographic images that typically belong to the same event, automatically select the most important image in terms of relative appeal, called emphasis, i.e., the one that should receive the most attention or special treatment. Potential EIS applications include automatic albuming [39], image retrieval from databases, and wireless imaging [38].

Experiments have demonstrated that the majority of strong positive attributes for EIS belong to the people and composition/subject categories [45]. The following features were selected for EIS, based on a compromise between importance and implementation cost: (a) features based on objective measures: sharpness, colorfulness, format uniqueness; (b) features related to people: skin area, people present, close-up; (c) features related to composition: main subject size, centrality, compactness, and variation. A brief description of these features is provided below.

A measure of colorfulness is obtained by examining for the presence of high-saturation colors along various hues. The chrominance plane is quantized into 12 bins and high saturation (above a threshold) pixels in each bin are counted to determine whether a bin is active. The colorfulness measure is determined based on the number of active bins:

$$\textit{Colorfulness} = \min\left\{\frac{\textit{Number of active bins}}{10}, 1.0\right\}. \quad (11)$$

Sharpness is estimated using the edge profile of the image. Image edges are detected from the smoothed luminance channel using the Sobel operator. The edge histogram is formed and the regions that contain the strongest edges are identified as those that are above the 90th percentile of the edge histogram. Strong-edge regions are refined via median filtering, and the average of the strongest edges provides an estimate of sharpness.

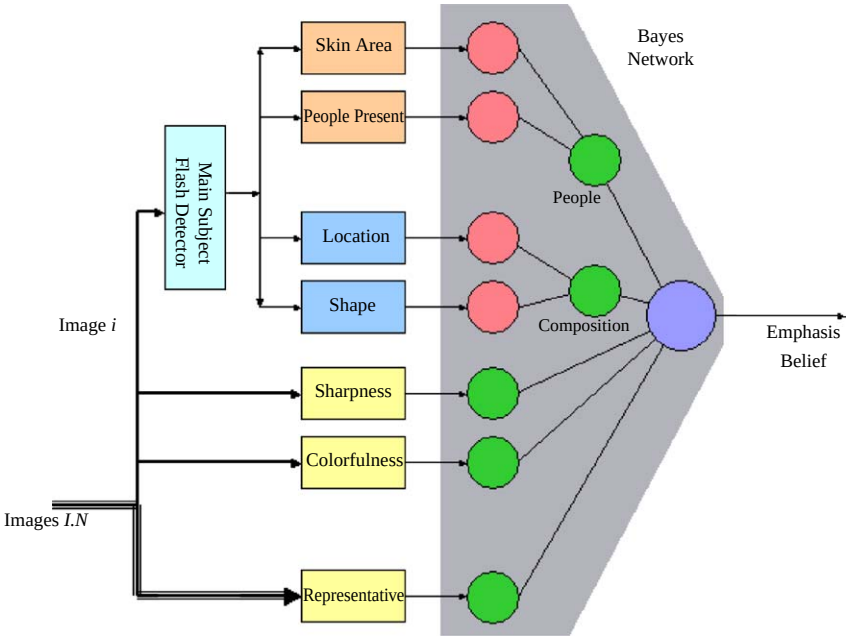

Fig. 4. A BN for emphasis value computation: (a) GT value = 58, (b) GT value = 70, (c) GT value = 93, (d) GT value = 94, (e) GT value = 54, (f) GT value = 65, (g) GT value = 93, (h) GT value = 95.

Format uniqueness is useful with APS pictures that have various aspect dimensions. When a picture is the only "Panoramic" picture in a group, it is more likely to be selected as the emphasis image.

Skin detection methods [20] enable the computation of people-related features. The percentage of skin/face area in a picture is computed as a preliminary step for determining the presence of people. The presence of people is detected when a significant amount of skin area is present in the image. The number of skin pixels in the image is counted and people are assumed present when the skin area is above a threshold. *People present* is a binary feature indicating the presence or absence of people. *Close-up* is determined as the percentage of skin area in the central portion of the image.

Good composition is the most important positive attribute of picture emphasis and bad composition is the most important negative attribute. However, composition is very difficult to determine through automatic means. The quantized main subject map is used to obtain an estimate of the image composition. Centrality is computed based on the location of the centroid of the most salient main subject regions. Compactness is an estimate of the concavity of the most salient main subject region. Variation depends on the intensity difference between the most salient main subject region and its surroundings.

Similar to MSD, a MLBN is used in the EIS system. The selected features are evaluated for dependencies and organized into sub-tree structures when the conditional independence assumption cannot be asserted in a single-level structure. For EIS, an intermediate "people" node is incorporated to alleviate the dependency between "Skin" and "People Present" nodes, and a "Composition" node is introduced to handle the features related to the main subject of an image. Each of the remaining features is considered independent of the other features and, therefore, is directly linked to the root node. The emphasis score is determined at the root node and all the feature detectors are at the leaf nodes. The outputs of the feature extraction stage represent statistical evidences that are integrated by the BN to compute the belief that the processed image is the emphasis image, as shown in Fig. 4. After the emphasis values have been computed for all images, the one with the highest emphasis is chosen as the emphasis image.

It should be noted that each link is assumed to be conditionally independent of other links at the same level, which results in convenient training of the entire net by training each link separately, i.e., deriving the CPM for a given link independent of others. In practice, this assumption is sometimes violated; however, the independence simplification makes implementation feasible and produces reasonable results.

The prototype EIS algorithm was trained based on 14 image groups for which one or two statistically separable emphasis images could be identified. It was then used to process all 194 images of the 30 image groups used in the ground truth experiment. Image examples are shown in Fig. 5. The results illustrated that the top ranked emphasis image was automatically selected in 40% of the cases, while an

Fig. 5. Example of EIS results for two groups of images with ground truth values; image (d) was selected as the emphasis image in groups (a)–(d); image (g) was selected as the emphasis image for group (e)–(h).

acceptable emphasis image was selected in 100% of the cases.

4.3. Image and scene classification (ISC)

Scene categorization is valuable in image retrieval from databases because an understanding of scene content can be used for efficient and effective database organization and browsing. In addition, image filtering and enhancement operations may be adjusted depending on scene type, so that the best rendering can be achieved. Knowledge of the scene type is also useful for event classification, which constitutes a fundamental component of automatic albuming systems [39]. The general problem of automatic scene categorization is difficult to solve and is best approached by a divide-and-conquer strategy. A good first step is to consider only two classes such as indoor vs. outdoor [1,40], which may be further subdivided into city vs. landscape [2], etc.

The problem of scene categorization is often approached by computing low-level features, e.g., color or texture [1], which are processed with a classifier engine for inferring high-level information about the image. One of the issues when dealing with a diverse set of features is how to integrate them into a classification engine. The solution proposed in Ref. [1] was to independently classify image subsections based on color and texture and obtain a final result using a majority classifier. An alternative method using SVM was introduced in Ref. [13].

One problem with the methods using low-level features in scene categorization is that it is often difficult to generalize these methods to diverse image data beyond the training set. More importantly, they lack semantic image interpretation that is extremely valuable in determining the scene type. Scene content such as the presence of people, sky, grass, etc., may be used as cues for improving the classification performance obtained by low-level features alone [19]. Sky and grass regions are identified using color and texture features and classifiers that are tuned for sky and grass detection (see details in Ref. [19]).

The BN structure, shown in Fig. 6, for classification of images to indoor vs. outdoor was constructed using the same procedure as described in the previous sections. The network integrates low-level features (color and texture) and semantic features (sky and grass) using a single classification engine. The network structure was designed according to domain knowledge and the general principles of BN. Recently, an automated learning algorithm confirmed the network structure [42]. This approach improves the classification performance over using low-level features alone. The CPMs for each node were derived using the frequency counting approach based on a Kodak database of consumer images [1]. The color features are based on the quantized color histogram (3×64 bins) in the Ohta color space [43], and texture features were based on the *Multiresolution Simultaneous Autoregressive* (MRSAR) model [1]. The classification based on color or texture was based on the k-nearest

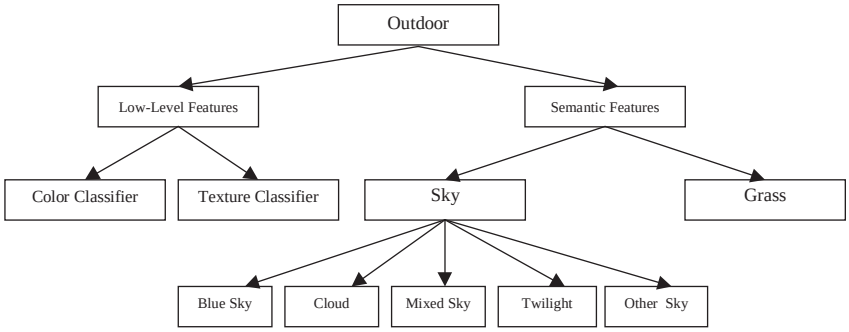

Fig. 6. A BN for indoor vs. outdoor scene classification.

Table 1
Indoor vs. outdoor classification results with integration of low-level and semantic features

Indoor vs. outdoor classification: percent correct results						
	C	T	C + T	C + S	T + S	C + T + S
Actual semantics	74.2%	82.2%	82.3%	80.9%	86.9%	90.1%
Computed semantics	Same	Same	Same	75.2%	84.0%	84.7%

Notation: C: color features; T: texture features; S: semantic sky and grass features.

neighbor classifier ($k = 1$), and yielded 74% and 82%, respectively for a database of 1300 images.

The sky and grass features were obtained using two methods. First the ground truth information about the images was used, i.e. the sky and grass detection is always correct. The indoor/outdoor classification results obtained this way reflect an upper bound, since the performance of any sky and grass classifier will be suboptimal. The second method involves using classification results to detect sky and grass information in the image. Sky and grass classification methods are based on color/texture features and yield a performance of 95% accuracy with 10% false positives.

The choice of threshold determines whether the image is indoor or outdoor. When the belief at the root node is above the threshold, the image is characterized as outdoor, thus, the network behaves as an outdoor detector. The threshold value was determined using one-fifth of the available data for training and the remaining data for testing. The value of 0.35 yielded the best overall results for both the training and testing data. In fact, the performance is statistically the same on both data sets.

The indoor/outdoor classification results with BN integration of low-level and semantic features are shown in Table 1. Semantic features are based on sky and grass information. The classification results are listed using actual semantics obtained from ground truth information, and computed semantics where sky and grass are detected using classification methods. In both cases, the use of semantic features improves the system performance and an overall accuracy of 90.1% (or 84.7%) was obtained when using both low-level and actual (or computed) semantic features. These results provide an improvement over the classification results based on the combination of color and texture (82.3%).

5. Benchmarking Bayes net performance

In this section, we describe a benchmarking study performed to compare the results of the BN-based automatic MSD system with other versions of the system built using one naïve and two different neural network-based classifiers. The naïve approach is implemented using a central zone predictor that divides the image into nine equal sized rectangular zones and marks the central zone as main subject and other zones as background. The neural network-based classifiers are similar to the BN-based classifiers in that they use a similar set of feature detectors to reason about the main subject regions. However, they differ from the BN in their training, implementation and knowledge propagation schemes. For this study, we have implemented two Bayesian network classifiers, one using a SLBN and the other using a MLBN, a naïve 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.

In each of the above cases, the MSD system provides results in the form of a belief map. Regions with highest belief

Table 2
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

represent higher likelihood of being the main subjects and those with low belief values represent background areas. Because of the uncertain nature of the problem, a simple correlation metric that counts the number of correctly classified and misclassified pixels cannot be used. We have developed an order-based correlation metric [46] derived from the Kemeny and Snell's distance [3] that produces similarity measurements more consistent with the subjective judgment of the third-party observers. We use this modified version of the Kemeny and Snell's distance metric (referred to as *dKS*) to evaluate the performance of each classifier on each image. This distance metric correlates the classifier output to the image ground truth in terms of the ordering of the various regions in the image. A smaller distance implies a more correlated ordering of the classifier output with the ground truth and, in turn, better classifier performance [21].

Once the *dKS* metric is applied to the output of the classifiers, a series of pair-wise direct comparisons can be conducted to analyze the performance of one classifier versus another. An analysis of these pair-wise comparisons in terms of the number of images on which each classifier performs better or worse than the other classifiers was performed. This analysis is highly dependent on the chosen metric and based on a small image set. To generate a more comprehensive picture of the performance of the various classifiers, we also perform various qualitative and statistical evaluations on the results. Subjective rankings show the opinion of three observers on the results of the various classification techniques. The analytical evaluations determine whether the differences observed between the results of the various classifiers have any statistical significance. We use the null hypotheses tests and the analysis of variance test on the various classifiers to compute statistical significance.

Table 2 shows the number of images on which the MLBN performs better than the other classification schemes. The multilevel Bayes net 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. Note that the central-zone, although simple, is not necessarily a poor predictor of main subject and, in fact, has been a standard used in many cameras for exposure and focus control. The improved performance over the SLBN is also expected as the multi-level network can make use of the full set of feature detectors whereas the single-level network uses only a subset based on the independency considerations between the variables. 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. Moreover, the full bank of features does not need to be computed for the BN, resulting in performance and speed scalable versions of the MSD [20].

In addition to the above comparison, two different statistical analysis tests were also performed on the data gathered from the *dKS* metric. The first of these was a series of null-hypothesis tests (also known as *t*-tests) that directly compared the performance of two classification schemes to determine if one was statistically significantly better than the other. The second test was the analysis of variance test (ANOVA) [44], which compared the performance of all the five classification schemes simultaneously. It produces the statistical significance relationships describing all the classifiers.

Fig. 7 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	

Fig. 7. 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. 8 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 naı¨ve predictor such as central zone.
2. The BN structure needs to be carefully constructed to account for dependencies between variables in the domain. 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.

6. Discussions and conclusions

In this paper, we 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, we demonstrated that the BN-based systems have excellent generalization on novel datasets. We 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

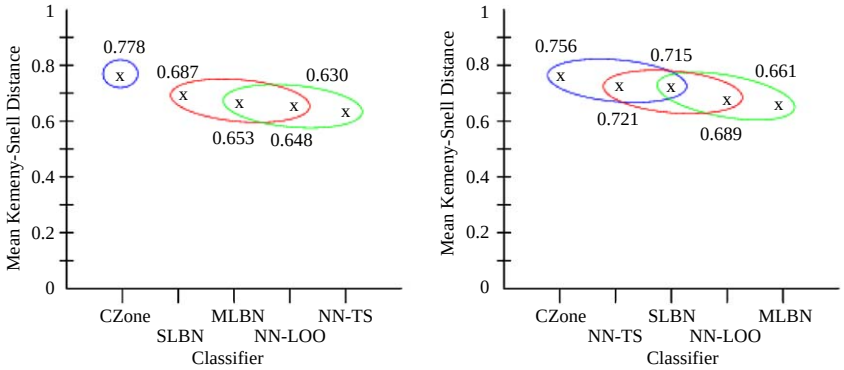

Fig. 8. Analysis of variance test on the *dKS* results of the five classifiers: Left—training set, Right—testing set.

memorize the training set, testing the BN-based systems generally does not give overly biased results.

Domain knowledge is critical for training a network that exhibits good generalization. A BN whose structure is determined based on domain knowledge is more likely to generalize well to novel data because it captures the underlying causal relationship among various cues. Fortunately, in semantic scene understanding of photographs, which contain subjects and content familiar to us, domain knowledge is not difficult to come by. In less obvious cases, psycho-visual studies such as those conducted in Refs. [36,45] are used.

Along the same line, the best approach to building BN may be to combine domain knowledge and automatic algorithms for discovering BN from data [34]. On one hand, relying on domain knowledge alone may miss certain causal relationships that are not obvious to even human experts; on the other hand, automatic algorithms may overfit the data by focusing on spurious correlation as opposed to true causality among cues. Using automatic algorithms and domain knowledge as a sanity check for each other can ensure a near-optimal network structure for a given application.

Semantic features are valuable, but they are not always present, or may not be detectable by the related algorithm even if they are present, in a given image. Therefore, it is critical to have an inference engine that is capable of handling missing cues properly and gracefully. BN are ideal for such cases, while discriminants such as neural networks, are poor in this regard. Compared to rule-based systems, which can be made to handle missing cues, BN provide a more principled alternative.

In conclusion, the general framework for semantic image understanding presented here can be applied to various tasks involving semantic understanding of pictorial images. With these diverse examples, we 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.

References

[1] M. Szummer, R.W. Picard, Indoor–outdoor image classification, Proceedings of Workshop on Content-based Access of Image and Video Database, 1998.

[2] A. Vailaya, A. Jain, H.J. Zhang, On image classification: city images vs. landscapes, Pattern Recognition (1998) 1921–1935.

[3] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. PAMI 22 (2000) 1349–1380.

[4] M. Kam, X. Zhu, P. Kalata, Sensor fusion for mobile robot navigation, Proc. IEEE 85 (1) (1997) 108–119.

[5] L.A. Klein, Sensor and Data Fusion Concepts and Applications, SPIE Press, Bellingham, WA, 1999.

[6] G.A. Baraghimian, A. Klinger, Preference voting for sensor fusion, Proceedings of the International Society of Optical Engineering, Aerospace Sensing, Sensor Fusion III, 1990, pp. 46–57.

[7] P. Pirjanian, H.I. Christensen, J.A. Fayman, Application of voting to fusion of purposive modules: an experimental investigation, Invited paper in J. Robotics Autonomous Systems 23 (4) (1998) 253–266.

[8] T. Brotherton, P. Simpson, Fuzzy neural networks for hierarchical fusion with applications, Proceedings of International Conference on Multisensor Fusion and Integration for Intelligent Systems, 1994.

[9] E. Jouseau, B. Dorizzi, Neural networks and fuzzy data fusion. Application to an online and real time vehicle detection system, Pattern Recognition Lett. 20 (1999) 97–107.

[10] L. Rong, Z. Wang, Fusion of numerical and linguistic information by the use of fuzzy neural network, Proceedings of Multisource-Multisensor Information Fusion, July 6–9, 1998.

[11] S.L. Lauritzen, Graphical Models, Oxford University Press, Oxford, 1996.

[12] Y. Wang, H.-J. Zhang, Content-based image orientation detection with support vector machine, Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries, 2001.

[13] N. Serrano, A. Savakis, J. Luo, A computationally efficient approach to indoor/outdoor scene classification, Proceedings of the International Conference on Pattern Recognition, 2002.

[15] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Francisco, 1988.

[16] A. Vailaya, M. Figueriredo, A. Jain, H.-J. Zhang, Content-based hierarchical classification of vacation images, Proceedings of IEEE International Conference on Multimedia Computing and Systems, June 1999.

[17] A. Vailaya, H.-J. Zhang, A.K. Jain, Automatic image orientation detection, Proceedings of IEEE International Conference on Image Processing, October 1999.

[18] S. Peak, S.-F. Chang, A knowledge engineering approach for image classification based on probabilistic reasoning systems, Proceedings of IEEE International Conference on Multimedia and Exposition, 2000.

[19] N. Serrano, A.E. Savakis, J. Luo, Improved scene classification using efficient low-level features and semantic cues, Pattern Recognition 37 (9) (2004) 1773–1784.

[20] J. Luo, S. Etz, A. Singhal, R.T. Gray, A computational approach to detecting main subjects in photographic images, Image Vision Comput. 22 (2004) 227–241.

[21] D. Ballard, C. Brown, Computer Vision, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[22] M. Sonka, V. Hlavac, R. Boyle, Image Processing, Analysis, and Machine Vision, second ed., Brooks & Cole Publishing, Pacific Grove, CA, 1999.

[23] D.E. Dudgeon, R.T. Lacoss, An overview of automatic target recognition, Lincoln Lab. J. 6 (1) (1993) 3–10.

[24] J. Schurmann, N. Bartneck, T. Bayer, J. Franke, E. Mandler, M. Oberlander, Document analysis—from pixels to contents, Proc. IEEE (1992) 1101–1119.

[25] G.P. Robinson, A.C.F. Colchester, Model-based recognition of anatomical objects from medical images, Image Vision Comput. 8 (1994) 499–507.

[26] P. Lipson, Context and configuration-based scene classification, Ph.D. Thesis, MIT, Cambridge, MA, 1996.
[27] P. Lipson, E. Grimson, P. Sinha, Configuration-based scene classification and image indexing, Proceedings of IEEE CVPR, 1997.
[28] J.R. Smith, C.-S. Li, Image classification and querying using composite region templates, Comput. Vision Image Understanding 75 (1/2) (1999) 165–174.
[29] M. Pradhan, et al., The sensitivity of belief networks to imprecise probabilities: an experimental investigation, Artif. Intell. 85 (1996).
[30] F.T. de Dombal, et al., Can computer aided teaching packages improve clinical care in patients with acute abdominal pain?, Br. Med. J. 302 (1991).
[31] C.K. Kwoh, D.F. Gillies, Using hidden nodes in Bayesian Network, Artif. Intell. 88 (1996).
[32] R.M. Fung, B. del Favero, Applying Bayesian Network to information retrieval, Commun. ACM 38 (3) (1995).
[33] S. Sarkar, S. Chavali, Modeling parameter space behavior of vision systems using BN, Comput. Vision Image Understanding 79 (2000).
[34] D. Heckerman, A tutorial on learning with Bayesian Network, Technical Report MSD-TR-95-06, Microsoft Research, March 1995.
[35] W. Buntine, Operations for learning with graphical models, Artif. Intell. Res. 2 (1994).
[36] S. Etz, J. Luo, Ground truth for training and evaluation of automatic main subject detection, Proceedings of Human Vision and Electronic Imaging, 2000.
[37] W. Osberger, A.J. Maeder, Automatic identification of perceptually important regions in an image, Proceedings of the International Conference on Image Processing, 1998.
[38] J. Luo, A. Singhal, A. Savakis, Efficient mobile imaging using emphasis image selection, Proceedings of the PICS Conference, 2003.
[39] A.C. Loui, A.E. Savakis, Automated Event Clustering and Quality Screening of Consumer Pictures for Digital Albuming, IEEE Trans. Multimedia, pp. 390–402, September 2003.
[40] S. Paek, C.L. Sable, V. Hatzivassiloglou, A. Jaimes, B.H. Schiffman, S.-F. Chang, K.R. McKeown, Integration of visual and text based approaches for the content labeling and classification of photographs, Proceedings of ACM SIGIR '99 Workshop on Multimedia Indexing and Retrieval, 1999.
[42] M.J. Kane, A.E. Savakis, Bayesian Network structure learning and inference in indoor vs. outdoor image classification, Proceedings of International Conference on Pattern Recognition, 2004.
[43] Y.-I. Ohta, T. Kanade, T. Sakai, Color information for region segmentation, Comput. Graphics Image Process. 13 (1980) 222–241.
[44] R.V. Hogg, E.A. Tanis, Probability and Statistical Inference, Macmillan, New York, 1988.
[45] A. Savakis, S.P. Etz, A.C. Loui, Evaluation of image appeal in consumer photography, Proceedings of Human Vision and Electronic Imaging, 2000.
[46] S. Etz, et al., Quantitative evaluation of rank-order similarity of images, Proceedings of IEEE International Conference on Image Processing, 2000.
[47] J. Antonisse, K.S. Keller, Dynamic evaluation of sources in rule-based sensor fusion systems, Proceedings of the Data Fusion Symposium, 1988.

About the Author—JIEBO LUO received his Ph.D. degree in Electrical Engineering from the University of Rochester in 1995. He is currently a Senior Principal Research Scientist in the Eastman Kodak Research Laboratories. His research interests include image processing, pattern recognition, and computer vision. He has authored over 80 technical papers and holds over 20 granted US patents. Dr. Luo was the Chair of the Rochester Section of the IEEE Signal Processing Society in 2001, and the General Co-Chair of the IEEE Western New York Workshop on Image Processing in 2000 and 2001. He was also a member of the Organizing Committee of the 2002 IEEE International Conference on Image Processing and a Guest Co-Editor for the Journal of Wireless Communications and Mobile Computing Special Issue on Multimedia Over Mobile IP. Currently, he is serving as an Associate Editor of the journal of Pattern Recognition and Journal of Electronic Imaging, an adjunct faculty member at Rochester Institute of Technology, and a member of the Kodak Research Scientific Council. Dr. Luo is a Senior Member of the IEEE.

About the Author—ANDREAS SAVAKIS received the BS (Summa Cum Laude) and MS degrees from Old Dominion University in 1984 and 1986, respectively, and the Ph.D. degree from North Carolina State University in 1991, all in Electrical Engineering. From 1991 to 1996 he conducted research in pattern recognition and visual memory as Research Associate at the Center for Electronic Imaging Systems at the University of Rochester and Research Assistant Professor at the University of Rochester Medical Center. From 1996 to 2000 he was with the Eastman Kodak Company, and held the positions of Research Scientist at the Business Imaging Systems Division and Senior Research Scientist at the Kodak Research Laboratories. In 2000, he joined the department of Computer Engineering at the Rochester Institute of Technology, where he is currently Professor and serves as Department Head. His research interests are in the areas of image understanding, multimedia applications, document image processing and hardware implementation of image processing and computer vision algorithms. Dr. Savakis is senior member of the IEEE, and member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi. He has served as Chair of the IEEE Western New York Image Processing Workshop, Chair of the Rochester Chapter of the IEEE Signal Processing Society, Treasurer of the IEEE Rochester Section and was presented with the IEEE Third Millennium Medal.

About the Author—AMIT SINGHAL received his Bachelor of Science degree, summa-cum-laude, in Computer Science and Engineering Technology from West Texas A&M University in May 1995. He then joined the Computer Science department at the University of Rochester for graduate studies, completing his Master of Science degree in May 1996 and earning his Ph.D. in Computer Science in August 2001. Amit joined the Imaging Science and Technology Laboratory at Kodak as a Senior Research Scientist in July 2000. There, he was responsible for

conducting leading research in image analysis and image understanding and for guiding research into Kodak products and technologies. He is currently a Principal Research Scientist with the Foundation Science Center at Eastman Kodak Company. His research interests include image understanding, image analysis, knowledge engineering, data fusion, and classification methodologies. He has worked in application areas as diverse as autonomous mobile robot navigation, digital photofinishing and digital imaging, and medical imaging. He also serves as a mentor to a doctoral Kodak Research Fellow. Amit has authored over 30 journal and conference papers and holds 1 US patent. He is a member of SPIE and ISIF.