Pattern Recognition 43 (2010) 4042–4054

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Semantic modeling of natural scenes based on contextual Bayesian networks

Huanhuan Cheng*, Runsheng Wang

*ATR National Laboratory, Institute of Electronic Science and Engineering, National University of Defense Technology, Changsha, Hunan 410073, China*

ARTICLE INFO

*Article history:*
Received 1 August 2009
Received in revised form
1 June 2010
Accepted 4 June 2010

*Keywords:*
Scene classification
Image representation
Bayesian network
Spatial information
Semantic features

ABSTRACT

This paper presents a novel approach based on contextual Bayesian networks (CBN) for natural scene modeling and classification. The structure of the CBN is derived based on domain knowledge, and parameters are learned from training images. For test images, the hybrid streams of semantic features of image content and spatial information are piped into the CBN-based inference engine, which is capable of incorporating domain knowledge as well as dealing with a number of input evidences, producing the category labels of the entire image. We demonstrate the promise of this approach for natural scene classification, comparing it with several state-of-art approaches.



## 1. Introduction

Scene classification, categorizing images into discrete categories (e.g., beach, forest or indoor), is a classical yet challenging problem in computer vision. It is an intermediate step to close the semantic gap between the image understanding of the user and the computer. From the application viewpoint, scene classification is relevant in systems for organization of personal and professional image and video collections. As such, this problem has been widely explored in the context of content-based image retrieval [1], but most prior approaches [2–4] have focused on mapping a set of classic low-level vision features to semantically meaningful categories using a classifier engine.

The semantic modeling of scenes by an intermediate representation was next proposed in order to reduce the gap between low-level and high-level image processing. The meaning of the semantic of the scene is not unique, and two basic strategies based on semantic representation can be found in literatures. One is the object-based strategy [5–8], which identifies the semantic as a set of materials or objects that appear in the image (e.g., sky, grass and rocks). These methods are mainly based on first segmenting the image in order to deal with different regions. Subsequently local classifiers are used to label the regions as belonging to an object. Finally, using this local information, the global scene is classified. Another popular approach is the bag-of-words strategy [9–13], which uses more general intermediate representations. In this case, they first identify a dictionary of visual words or local semantic concepts in order to build the bag-of-words, and further use bag-of-words models (e.g., probabilistic latent semantic analysis (pLSA) [14] and latent Dirichlet allocation (LDA) [15]) to discover clusters of local semantic concepts for scenes.

Although these two strategies have both achieved some promising results, images with similar visual contents are often mis-categorized, particularly for natural scenes. For example, in experiments on Vogel's dataset of natural scenes [7], coasts and river/lakes are frequently confused, and the reported performance of river/lakes is less than the average rate. This is also observed in the result based on bag-of-words methods [12,16]. Fig. 1 shows two images from Vogel's dataset. For the first river/lakes scene, the percentages of water, sky and rocks are very similar with those in coasts images. The second image is clearly a forest scene. However, the large amount of grass causes the image high probability to be classified as a plain scene. In other words, common materials of scenes usually produce similar semantic features. For this reason, even the approaches based on semantic modeling fail to distinguish them correctly. Therefore, this shows that there is still a challenging work, and more advanced classification methods need to be designed for scene classification.

For natural scenes, a scene is generally composed of several entities, organized in often unpredictable layouts, varying with different seasons and weathers. This makes natural scenes hard to be distinguished. However, without any accurate features extracted from images, people can categorize images into natural scenes very well. Human perception mainly relies on domain knowledge about certain scenes, which includes various attributes such as objects' occurrence probabilities and their spatial

* Corresponding author. Tel.: +86 731 4575724; fax: 86 731 4518730.
*E-mail address:* ch2huan@nudt.edu.cn (H. Cheng).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.06.004

River/lake Forest

**Fig. 1.** Examples of images which are often misclassified.

arrangements. Furthermore, these various elements will be considered in a unified way when people categorize a picture. The challenge with such an idea 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. Thus, in the paper, we aim to develop a unified framework to model and classify images by various attributes.

Bayesian networks (BN) provide a powerful framework for knowledge representation and domain-specific knowledge can be incorporated in the network structure. Luo et al. [5] present a general-purpose knowledge integration framework for semantic image understanding that employs BN in integrating both low-level and semantic features. Indoor/outdoor classification is one of their applications. In their work, semantic features are few and spatial information is not taken into account.

We believe that context cues in images should be used for a more intuitive and accurate classification. The spatial correlation of basic elements (e.g., pixels, lines and regions) is the most common type of context in the image. While, in habitual bag-of-words techniques, the spatial relationships of the image patches or object parts are ignored. The object-based strategy can explicitly exploit spatial information among the parts or regions. Therefore, we follow the object-based strategy in this paper.

In this paper, we propose a contextual Bayesian network-based framework for natural scene modeling and classification, which models spatial relationships between local semantic regions and integrates multi-attributes to infer high-level semantic of a global image. The contributions of our paper are the following:

(1) A unified probabilistic framework for scene classification based on Bayesian networks to represent scenes by various attributes. We propose a contextual Bayesian network to model local materials of images and spatial configurations of scenes together, which have been proven to classify natural scenes successfully. This is different from previous works such as [5,8], where only two scene categories (indoor/outdoor) are considered and the spatial relationships are ignored.
(2) An effective approach based on the use of spatial information of the key entities for scene classification. This spatial information is not fixed, varying with key semantic regions chosen in different images. Besides, it allows for a simpler representation for scene structure, which has proven to be helpful to categorize some type of natural scenes, which are often mis-categorized.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 describes the general framework we explore. The contextual Bayesian network model is presented in Section 4. Classification results are provided and discussed in Section 5. Section 6 concludes the paper.

## 2. Related work

Early works on scene classification use low-level features directly from the whole image or from a fixed spatial layout, combining with supervised learning methods to classify images into several semantic classes. The work by Vailaya et al. [2] is regarded as a representative of the literature in this field. This approach relies on a combination of distinct low-level cues for different two-class problems (global edge features for city/landscape and local color features for indoor/outdoor).

The semantic modeling of scene classification can be primarily categorized into bag-of-words methods and object-based methods. In recent years, bag-of-words models have shown much success for text analysis and information retrieval. Inspired by this, a number of works [9–12] propose the demonstrated impressive results for image analysis and classification using the bag-of-words models. Bosch et al. [9] provide an approach, which uses bag-of-words to model visual scenes based on local invariant features and probabilistic latent semantic analysis (pLSA). The same authors extend their work to investigate the various choice of vocabularies, parameters and the gain in adding spatial information [10]. Fei-Fei and Perona [11] independently propose two variations of LDA. In that framework, local regions are first clustered into different intermediate themes, and then into categories. No supervision is needed apart from a single category label to the training image.

Several studies suggest that to understand the context of a complex scene, one needs first to recognize the objects and then in turn recognize the category of the scene [17]. The object-based methods are following this strategy and are closer to human perceptions. Luo et al. [5] proposed a hybrid approach: low-level and semantic features are integrated into a general-purpose knowledge framework that employs a Bayesian network (BN). Vogel and Schiele [7] recently present a novel image representation for natural scene modeling by local semantic description. They predefined a set of semantic concepts such as water, rocks and foliage to describe the content of images. They first classify local image regions into semantic concept classes. Images are represented through the frequency of occurrence of these local concepts. But spatial relationships between objects are not considered in these works.

Aksoy et al. [18] applied a Bayesian framework in a visual grammar. Scene representation is achieved by decomposing the image into prototype regions and modeling the interactions between these regions in terms of their spatial relationships. Boutell et al. [19] present a graph-based approach to learn spatial configuration models for outdoor scenes. Since a fully connected scene configuration model is intractable, they latterly chose to model pairwise relationships between regions and estimate scene probabilities using loopy belief propagation on a factor graph in [20]. This generative model offers a number of advantages at the expense of slightly lower accuracies compared with discriminative models using same semantic features.

These object-based approaches are able to provide a visual representation of objects based on image regions. However, many of them lack sufficient use of spatial context information and an effective mapping mechanism from diverse feature set to high-level semantic features contained in global pictures.

Therefore, this paper proposes a contextual Bayesian network framework to categorize natural scenes. The structure of the CBN is derived based on domain knowledge, and parameters are learned from training images. For test images, they are first segmented into homogeneous regions, and labeled by the local classifier with an object by their identities. Then the semantic features and the spatial relationships of key semantic regions are extracted from images. The hybrid stream of these