Multimed Tools Appl (2015) 74:4965–4986
DOI 10.1007/s11042-014-1858-9

Bayesian network based semantic image classification with attributed relational graph

Chang-Yong Ri · Min Yao

Published online: 7 February 2014

Abstract Semantic image classification is a hot issue of image mining. Information of spatial relations between objects in an image is one of the important semantic information of an image. However, the previous researches have not made full use of the spatial relations for image modeling and classification. In addition, to classify the images with Bayesian network, the accuracy of conditional probability estimation may be insufficient, because the learning methods of spatial contextual models have usually used a limited number of training samples. In this work, the semantic image modeling based on attributed relational graph has been proposed, in which the distance measure method between images was presented, therefore the object information and spatial relational information could be fully utilized. Then, the semantic distance between images based on attributed relational graph could be calculated for the support vector machine to obtain the joint conditional probability distribution of Bayesian network. Therefore the probabilistic estimation problem under the sparse training samples could be solved, and the accuracy of semantic image classification with Bayesian network was improved. Experimental results show the validity and reliability of this proposed method.

Keywords Semantic image classification · Attributed relational graph · Semantic distance · Bayesian network

1 Introduction

A natural scene image is composed of several entities and their features are affected by weather, season, camera position etc., therefore its classification by computer is difficult. However, human being correctly recognizes and classifies them through domain knowledge about certain scenes, which includes the objects' presences and their contextual information.

CY. Ri (✉) · M. Yao
School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
e-mail: changyongri@gmail.com

M. Yao
e-mail: myao@zju.edu.cn

CY. Ri
Institute of Information Science, Kim Il Song University, Pyongyang 190016,
Democratic People's Republic of Korea

The contextual information of an image is important to help the semantic image understanding and classification [15, 29, 33]. Therefore, in order to improve the accuracy of image recognition and classification, many researchers have attempted to integrate the contextual information into their recognition systems [1, 6, 9, 11, 16, 26]. They have described an image with the objects and the relations between objects in an image: visual grammar model based on Bayesian framework [1], region-based scene configuration model [6], contextual Bayesian network (CBN) model [11] etc. However, most of researchers have only used the predefined and fixed spatial relations or have not made full use of the spatial relations for image modeling and classification. Therefore, it is challenging and significant how to reflect the contextual relations in image modeling sufficiently and to apply the modeling to image classification. This is the first motivation for the approach to semantic image representation and classification proposed in the rest of this presentation. The other motivation is that an image description based on graphical models accords with the human perception of image knowledge, i.e. the semantic objects and spatial arrangements in an image [6, 11, 13, 17]. Semantic information of an image includes the semantic objects contained in an image and the relations between objects.

Concretely, we propose an image description model based on attributed relational graph (ARG), thus make full use of the spatial relations between objects in an image. ARG [4] is one of the graphical models to solve the problems of the computer vision and image processing fields. In our image description model, an image is represented by graph of which nodes and edges correspond to the semantic objects and their spatial relations. The nodes and edges have their attributes which reflect the semantic information such as object occurrence and spatial arrangements of an image. In addition, we propose a semantic distance measure method between images based on ARG model. Compared with the existing graphical models [13, 17], we use not regular image patch but segmented semantic objects and investigate the spatial arrangements in an image more sufficiently by defining fuzzy contextual relations. Moreover, the proposed semantic distance measure is used in image classification, where we exploit the Bayesian network. Bayesian network provide a powerful framework for integrating of the various information and have been successfully used in image classification [11, 22, 27]. However, the learning methods in Bayesian network usually train with limited number of samples, thence the accuracy of obtained conditional probabilities may be insufficient. In this paper, we obtain the joint conditional probability by support vector machine (SVM) where the proposed distance measure is used to construct the kernel function, thus ensure the accuracy of probability distribution with sparse training samples. Through the Bayesian network, we achieve the integration of the semantic object information and spatial arrangement information of an image. The contributions of our paper are as follows:

1) We propose an image description model in the form of attributed relational graphs and semantic distance measure method with this model. This can effectively reflect the semantic object information and spatial arrangement information of images. This is different from previous graphical models such as Olivier et al. [13], Harchaoui et al. [17] and Cheng et al. [11]: where they have considered spatial relations through adjacent regions (regular grid regions [13] or segments [17]) merely or key object regions [11], thus not fully reflect the spatial arrangement of an image.
2) We obtain the joint conditional probability of Bayesian network for image classification by SVM based on proposed distance measure. Compared with the state of the art for image classification [11, 27], this can ensure the accuracy of probability distribution with sparse training samples and improve the classification accuracy through the proposed method of semantic distance measure between images.

1.1 Related work

Image classification which categorizes images into discrete classes is a challenging problem in computer vision, and has attracted considerable attention. Many early researches have attempted to map the low-level visual features such as color, texture, etc. to high-level semantic categories (e.g., beach, forest or indoor) [10, 28, 31]. Some researches [3, 23] have described image content with MPEG-7 visual descriptors, namely *dominant color*, *homogeneous texture*, *region shape* and *edge histogram*. In Papadopoulos et al. [23], MPEG-7 descriptors were extracted and concatenated to form the region feature vector, and the different classifiers were employed to semantic region annotation. Athanasiadis et al. [3] used MPEG-7 visual descriptors to characterize a region in terms of low-level features. They also used three types of relations (i.e. *specialization relation*, *part of relation* and *property relation*), of which semantics are defined in the MPEG-7 standard, to describe the spatial relations between regions. However, MPEG-7 visual descriptors are the low-level visual features essentially. In order to reduce the semantic gap between low-level visual features and high-level semantic information, the semantic modeling by an intermediate representation of an image has next proposed, where the bag-of-words (BoW) [5, 18, 21, 30, 34] models and the object (or region)-based models [11, 20, 32] have been dominate.

In the BoW models, the local features extracted from images are first mapped to a set of visual words, and the images are described as a bag of discrete visual words, then the frequency distributions of these words are used for image categorization. Fei-Fei et al. [14] presented an image description model by a collection of local image regions which are denoted as code words obtained by unsupervised learning. Liu et al. [21] presented a method for image classification by integrating region contextual information into the BoW approach. In contrast to the traditional BoW approach which learned each visual word independently according to its visual feature, they append the region contextual constraint into this framework. Since the constituent patches of a region do not exist in the isolation, the region contextual information can well capture intrinsic property of patches. Under the BoW framework, Yang et al. [34] presented an object representation model that incorporates the object appearance and contextual information from multiple spatial levels for robust object categorization. Bosch et al. [5] also employed the BoWs to model visual scenes based on local invariant features and probabilistic latent semantic analysis (pLSA). In order to improve the performance of the BoW model for image classification, Su et al. [30] have combined bag-of-words histograms with semantic image descriptors at decision level and integrated the semantic information into the visual vocabulary.

Compared to the BoW models, the object-based models identify the semantic concept as a set of materials or objects that appear in image (e.g., sky, grass, building, cars etc.), and then develop generative models of the images. In the object-based models, firstly the object recognition is performed and then the scene categories are classified. Luo et al. [22] presented a semantic-based image description framework that employs Bayesian network in integrating low-level features and semantic information, and therefore an image has been classified into indoor or outdoor scenes. For the purpose of the semantic-based image description, Vogel et al. [32] proposed a model based on concept occurrence vector (COV), where the local image patches have been corresponded to semantic object concepts and images have been described with the frequency of occurrence of the semantic concepts. They carried out the image classification with the category prototype method and the SVM-based method. Cheng et al. [11] proposed the contextual Bayesian network (CBN) model for natural scene modeling and classification, where the hybrid streams of object occurrence information and spatial arrangement information of image are piped into the CBN-based inference engine. In their work, the

images have been manually or automatically segmented and annotated as the semantic objects, and then the spatial relations between objects are computed through the key objects. Some researches showed that the object-based methods are more close to human perceptions [11] and often outperform the BoWs in terms of classification rates [8]. Therefore, we handle our classification task of image on the object-based strategy.

As it is mentioned above, the contextual information of an image is important to help the semantic image understanding and classification. And there are some researches to improve the performance of image classification by introducing the contextual information of an image [1, 6, 9, 11, 12, 16, 19, 21, 26, 30, 34]. In Papadopoulos et al. [23], the fuzzy directional relations between image regions have been considered as contextual information, and the object-level spatial contextual information has been used to classify the images. Bruzzoneet al. [7] proposed a context-sensitive semi-supervised SVM classifier, where the contextual information of the adjacent pixels is introduced into the cost function of the classifier. Considering the inner-class difference of an image category, Qi et al. [25] proposed a method to construct a prototype set of the spatial contextual models by employing the kernel methods, and carried out the image classification using the distance measure between image models. However, these works have only used the predefined and fixed spatial relations or have not made full use of the spatial relations for image modeling and classification. In recent years, some researchers attempt to describe the images with graphical models which have a good ability to describe the semantic image contents [6, 11, 13, 17, 24]. Olivier et al. [13] proposed to represent images by graphs of which nodes represent the image grid regions, and edges represent the adjacency relationships. They have defined the matching between two images as the optimization of energy in multi-label Markov random fields (MRF) which are defined on the corresponding graphs. Then they considered the value of the optimized MRF associated with two images as a kernel, performed image classification by SVM classifier. Harchaoui et al. [17] also employed graphical representation of an image, and proposed a family of kernels based on sub-tree patterns of the graphs, in turn employed the semi-supervised learning method and the multiple-kernel learning method for image classification. However, they have also considered spatial relations through adjacent regions (regular grid regions or segments) merely, thus not fully reflect the spatial arrangement of an image.

It is worth mentioning that most of the studies on image classification has employed SVM [7, 19, 32] or Bayesian classifiers [1, 14, 22, 27]. SVM classifiers have advantages in solving classification problems with sparse samples, nonlinear and high dimensional data. In contrast, Bayesian networks provide a powerful framework for knowledge representation, i.e. have advantages in integrating different information. However, the learning methods in Bayesian network usually take the limited training samples, therefore the accuracy of estimation of conditional probabilities may be insufficient. Some researchers have employed SVM and Bayesian classifiers together for their image classification tasks [11, 27]. Inspired from their researches, we can obtain the conditional probability by SVM classifier, according to its ability of classification, and achieve the integration of different information of an image by Bayesian network.

In conclusion, we found that the above mentioned works of semantic image classification using contextual information have two defects: they have only used the predefined and fixed spatial relations or have not made full use of the spatial relations, and the learning methods of spatial contextual models have usually used sparse training samples which can lead to the insufficient accuracy of obtained conditional probabilities. To address these disadvantages, we propose an image description model based on attributed relational graph which make full use of the spatial relations between objects in an image by defining fuzzy contextual relations. Afterword, we propose a semantic distance measure method between attributed relational

graphs, and obtain the joint conditional probability by SVM, thus ensure the accuracy of probability distribution with sparse training samples. Finally, we carry out the natural scene image classification using Bayesian network, where the semantic object information and spatial structure information of an image are integrated. Therefore we improve the accuracy of image classification. Experiments on four benchmark image databases (image dataset provided by Vogel and Schiele [32], image dataset provided by Fei-Fei and Perona [14], LabelMe and Caltech-101) show that the proposed methods can improve the performance of the natural scene classification compared with the state of the art results [6, 11, 13, 19, 21, 32]. It shows the validity and suitability of these proposed methods.

2 Semantic image description based on attributed relational graph

2.1 Image description model based on attributed relational graph

Definition 1 An image can be modeled by following quadruple attributed relational graph:

$$I =< V, E, a, w > \tag{1}$$

where, V is a set of nodes in the graph I and it denotes a collection of semantic objects in the corresponding image; E is a set of edges in the graph I and it denotes a collection of spatial relations between objects in the corresponding image, $V \times V \rightarrow E$; a is a set of values of node's attribute, namely it is a set of values of the semantic object features in an image, $F^A: V \rightarrow a$; w is a set of value of edge's attributes, namely it is a set of value of the spatial relation features, $F^W: E \rightarrow w$.

In this work, the size of an object in image is considered as the attribute of corresponding node, and the directional, distance and topological relations between objects are considered as the attributes of corresponding edges.

The kind of semantic objects contained in an image of specific field is usually limited. For example, the natural scene images can be described by nine semantic objects, i.e. A = {sky, water, grass, trunks, foliage, field, rocks, flowers, sand}. With these nine semantic concepts, the natural scene images can be annotated to 99.5 %[32].

Definition 2 The value of attribute $a_i (a_i \in a)$ of a semantic object $v_i (v_i \in V)$ can be defined as the visual intensity of the object in an image, namely the percentage of the object's area in the whole image:

$$a_i = F^A(v_i) = \frac{|reg(v_i)|}{|I|} \tag{2}$$

where, $|reg(\cdot)|$ denotes the number of pixels of an object region; $|I|$ denotes the total number of pixels of an image.

The spatial relations between two objects can be divided into three classes: directional relations W_1, distance relations W_2 and topological relations W_3. The directional relations include *above*, *below*, *left* and *right*; the distance relations include *near* and *far*; the topological relations include *disjointed*, *bordering*, *invaded by* and *surrounded by* [2]. Moreover, these spatial relations can be combined into several classes, because the spatial relation between two object regions can be described by overlapping multiple relations, e.g. *invaded by* from *left*, *right* and *near*, etc.

Considering the characteristics of the natural scene images, i.e. *left* and *right* don't affect the image classification, the directional relations can be divided into *above*, *below* and *beside*. Moreover, *near* and *far* are inverse each other, namely, if the value of degree of *near* becomes higher, then the value of *far* becomes lower. Therefore, the distance relations can be described by only *near* (or *far*) enough. Similarly, the topological relations can be described by only *surrounded by*. It means *disjointed* and *bordering* that the value of degree of *surrounded by* is 0. If the value of degree of *surrounded by* is greater than 0, then the topological relation becomes *invaded by* or *surrounded by*. Especially, if the value of degree of *surrounded by* is 1, then the topological relation is the complete *surrounded by*. Figure 1 shows the spatial relations between objects in an image.

Definition 3 The value of attribute $w_{ij}(w_{ij} \in w)$ of a spatial relation $e_{ij}(e_{ij} \in E)$ between objects v_i and $v_j(v_i, v_j \in V)$ can be defined by the spatial relation descriptors $(\theta_{ij}, d_{ij}, \rho_{ij})$:

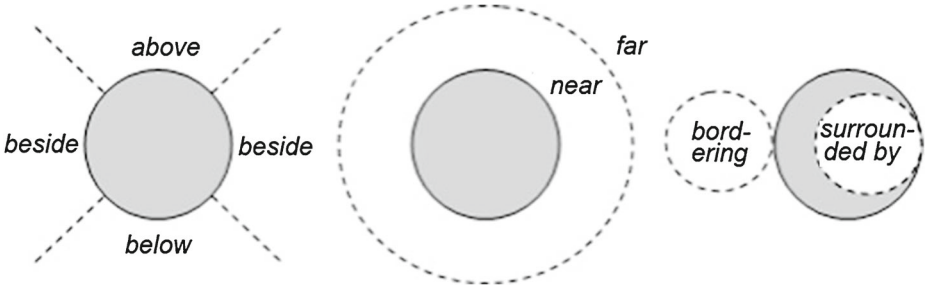
$$w_{ij} = \left(w_{1,ij}, w_{2,ij}, w_{3,ij}\right) = \left(F^{W_1}\left(e_{ij}\right), F^{W_2}\left(e_{ij}\right), F^{W_3}\left(e_{ij}\right)\right) = \left(\mu_1\left(\theta_{ij}\right), \mu_2\left(d_{ij}\right), \mu_3\left(\rho_{ij}\right)\right) \quad (3)$$

where, θ_{ij} denotes a angle between the horizontal axis and the line joining the centers of two object regions; d_{ij} denotes minimum distance between the boundary pixels of two object regions; ρ_{ij} denotes a ratio of the common perimeter between two object regions to the perimeter of the first object region.

In fact, the spatial relations between objects are the fuzzy relations. In Eq. (3), $\mu_R(R=1,2,3)$ are the fuzzy membership degrees which denote the belonging degrees of the spatial relations between objects to the directional relations, distance relations and topological relations, respectively. The membership degrees of the five fuzzy spatial relations between objects can be computed with spatial relation descriptors as follows.

Definition 4 The membership degrees of fuzzy spatial relations between objects in an image can be computed with the following equations:

$$\mu_{ABOVE}\left(\theta_{ij}\right) = \begin{cases} \sin^2\theta_{ij}, & if\ 0 < \theta_{ij} < \pi \\ 0, & otherwise \end{cases} \quad (4)$$



(a) Directional relations (b) Distance relations (c) Topological relations

Fig. 1 Spatial relations between objects in an image. **a** Directional relations. **b** Distance relations. **c** Topological relations

$$\mu_{BELOW}(\theta_{ij}) = \begin{cases} \sin^2\theta_{ij}, & if\ -\pi < \theta_{ij} < 0 \\ 0, & otherwise \end{cases} \tag{5}$$

$$\mu_{BESIDE}(\theta_{ij}) = \cos^2\theta_{ij} \tag{6}$$

$$\mu_{NEAR}(d_{ij}) = \frac{1}{1 + e^{\alpha_1(d_{ij} - \beta_1)}} \tag{7}$$

$$\mu_{SUR}(\rho_{ij}) = \frac{1}{1 + e^{-\alpha_2(\rho_{ij} - \beta_2)}} \tag{8}$$

where, α_1 and α_2 are the parameters that determine the crispness of the fuzzy membership degrees for distance relations and topological relations, respectively; β_1 is the cut-off value that divides the distance relations into *near* and *far* fuzzy relations; β_2 is the cut-off value that determines the *surrounded by* fuzzy relations.

Finally, the concrete directional relations between objects can be determined by maximum membership principle:

$$W_{1,ij} = L^{W_1}(e_{ij}) = \underset{W \in \{ABOVE, BELOW, BESIDE\}}{\arg\max} \mu_W(\theta_{ij}) \tag{9}$$

where, $L^{W_1} : E \rightarrow W_1$ represent the mapping from the edges of the graph I to the labels of the directional relations.

2.2 Semantic distance measure between attributed relational graphs

The distance between attributed relational graphs should reflect the mismatching degree between semantic information of images. Semantic information of an image includes the semantic objects contained in image and the relations between objects, therefore, the distance between attributed relational graphs can be composed of two components: distance between sets of node's attributes and distance between sets of edge's attributes. In order to define the distance measure, firstly, define the "correspondence" between attributed relational graphs.

Definition 5 Correspondence between attributed relational graphs $I^{(1)}=<V^{(1)},E^{(1)},a^{(1)},w^{(1)}>$ and $I^{(2)}=<V^{(2)},E^{(2)},a^{(2)},w^{(2)}>$ can be expressed as $l: I^{(1)} \rightarrow I^{(2)}$ and defined as follows:

- *Correspondence between nodes*: each node $v_i^{(1)}$ of graph $I^{(1)}$ corresponds only to the unique node $v_i^{(2)}$ (it is expressed as $l(v_i^{(1)})=v_i^{(2)}$) or does not correspond to any node (it is expressed as $l(v_i^{(1)})=\phi$) of graph $I^{(2)}$; if $\forall v_i^{(1)}$, $v_j^{(1)} \in V^{(1)}$, $v_i^{(1)} \neq v_j^{(1)}$, $l(v_i^{(1)}) \neq \phi$, $l(v_j^{(1)}) \neq \phi$, then $l(v_i^{(1)}) \neq l(v_j^{(1)})$.
- *Correspondence between edges*: for each edge of graph $I^{(1)}$, $\forall e_{ij}^{(1)} \in E^{(1)}$, where the two nodes of edge $e_{ij}^{(1)}$ are $v_i^{(1)}, v_j^{(1)} \in V^{(1)}$, if $l(v_i^{(1)})=v_i^{(2)} \neq \phi$, $l(v_j^{(1)})=v_j^{(2)} \neq \phi$, and the edge between the nodes $v_i^{(2)}$ and $v_j^{(2)}$ is $e_{ij}^{(2)} \in E^{(2)}$ of graph $I^{(2)}$, then the edge $e_{ij}^{(1)}$ corresponds to the edge $e_{ij}^{(2)}$ (it is expressed as $l(e_{ij}^{(1)})=e_{ij}^{(2)}$), else $e_{ij}^{(1)}$ does not correspond to any edge of graph $I^{(2)}$ (it is expressed as $l(e_{ij}^{(1)})=\phi$).

In fact, a set of node's attributes and a set of edge's attributes are represented as vectors in feature spaces. Therefore, the distance between sets of attributes can be computed with Euclidean distance.

Definition 6 The distance between sets of node's attributes of $I^{(1)}$ and $I^{(2)}$ under the correspondence l can be defined as follows:

$$d_V\left(I^{(1)},I^{(2)}\middle|l\right)=\left[\sum_{v_i^{(1)}\in V^{(1)}}\left|F^A\left(v_i^{(1)}\right)-\lambda_1\cdot F^A\left(l\left(v_i^{(1)}\right)\right)\right|^2\right]^{1/2}$$
$$=\left[\sum_{\substack{v_i^{(1)}\in V^{(1)},\,v_i^{(2)}\in V^{(2)}\\ v_i^{(2)}=l\left(v_i^{(1)}\right)}}\left|F^A\left(v_i^{(1)}\right)-\lambda_1\cdot F^A\left(v_i^{(2)}\right)\right|^2\right]^{1/2}=\left[\sum_{\substack{v_i^{(1)}\in V^{(1)},\,v_i^{(2)}\in V^{(2)}\\ v_i^{(2)}=l\left(v_i^{(1)}\right)}}\left(a_i^{(1)}-\lambda_1\cdot a_i^{(2)}\right)^2\right]^{1/2} \quad (10)$$

where, $\lambda_1=\begin{cases}1, & \text{if } L^A\left(v_i^{(1)}\right)=L^A\left(l\left(v_i^{(1)}\right)\right)\\ 0, & \text{otherwise}\end{cases}$, $L^A: V\to A$ represent the mapping from the nodes of the graph I to the semantic concepts of objects.

Definition 7 The distance between sets of edge's attributes of $I^{(1)}$ and $I^{(2)}$ under the correspondence l can be defined as follows:

$$d_E\left(I^{(1)},I^{(2)}\middle|l\right)=\left[\sum_{e_{ij}^{(1)}\in E^{(1)}}\left|F^W\left(e_{ij}^{(1)}\right)-\lambda_2\cdot F^W\left(l\left(e_{ij}^{(1)}\right)\right)\right|^2\right]^{1/2}=\left[\sum_{\substack{e_{ij}^{(1)}\in E^{(1)},\ e_{ij}^{(2)}\in E^{(2)}\\ e_{ij}^2=l\left(e_{ij}^1\right)}}\left|F^W\left(e_{ij}^{(1)}\right)-\lambda_2\cdot F^W\left(e_{ij}^{(2)}\right)\right|^2\right]^{1/2}$$
$$=\left[\sum_{\substack{e_{ij}^1\in E^1,\ e_{ij}^2\in E^2\\ e_{ij}^{(2)}=l\left(e_{ij}^{(1)}\right)}}\left|w_{ij}^{(1)}-\lambda_2\cdot w_{ij}^{(2)}\right|^2\right]^{1/2}=\left[\sum_{\substack{e_{ij}^{(1)}\in E^{(1)},\ e_{ij}^{(2)}\in E^{(2)}\\ e_{ij}^{(2)}=l\left(e_{ij}^{(1)}\right)}}\left(\begin{array}{l}\left(w_{1,ij}^{(1)}-\lambda_2\cdot\lambda_3\cdot w_{1,ij}^{(2)}\right)^2+\left(w_{2,ij}^{(1)}-\lambda_2\cdot w_{2,ij}^{(2)}\right)^2\\ +\left(w_{3,ij}^{(1)}-\lambda_2\cdot w_{3,ij}^{(2)}\right)^2\end{array}\right)\right]^{1/2} \quad (11)$$

where, $\lambda_2=\begin{cases}1, & \text{if } L^A\left(v_i^{(1)}\right)=L^A\left(l\left(v_i^{(1)}\right)\right) \text{ and } L^A\left(v_j^{(1)}\right)=L^A\left(l\left(v_j^{(1)}\right)\right)\\ 0, & \text{otherwise}\end{cases}$; $\lambda_3=\begin{cases}1, & \text{if } L^{W_1}\left(e_{ij}^{(1)}\right)=L^{W_1}\left(l\left(e_{ij}^{(1)}\right)\right)\\ 0, & \text{otherwise}\end{cases}$.

Definition 8 The distance between the attributed relational graphs $I^{(1)}$ and $I^{(2)}$ can be defined as a minimum value of the distances under all possible correspondences $l_k: I^{(1)}\to I^{(2)}$ between $I^{(1)}$ and $I^{(2)}$:

$$d_t\left(I^{(1)},I^{(2)}\right)=\min_{l_k} d_t\left(I^{(1)},I^{(2)}\middle|l_k\right),\quad t\in\{V,E\} \quad (12)$$

If the correspondence l^* between $I^{(1)}$ and $I^{(2)}$ meet the following conditions:

- $\forall v_i^{(1)}\in V^{(1)},\exists v_i^{(2)}\in V^{(2)}, L^A(v_i^{(1)})=L^A(v_i^{(2)})\Rightarrow l^*(v_i^{(1)})=v_i^{(2)}$, or $\forall v_i^{(2)}\in V^{(2)}$, $L^A(v_i^{(1)})\neq L^A(v_i^{(2)})\Rightarrow l^*(v_i^{(1)})=\phi$
- $\forall e_{ij}^{(1)}\in E^{(1)}$, $\exists e_{ij}^{(2)}\in E^{(2)}$, $L^A\left(v_i^{(1)}\right)=L^A\left(v_i^{(2)}\right)$, $L^A\left(v_i^{(1)}\right)=L^A\left(v_j^{(2)}\right)$, $L^{W_1}\left(e_{ij}^{(1)}\right)=L^{W_1}\left(e_{ij}^{(2)}\right)\Rightarrow l^*\left(e_{ij}^{(1)}\right)=e_{ij}^{(2)}$, or $\forall e_{ij}^{(2)}\in E^{(2)}$, $\left(L^A\left(v_i^{(1)}\right)\neq L^A\left(v_i^{(2)}\right), \text{or } L^A\left(v_j^{(1)}\right)\neq L^A\left(v_j^{(2)}\right), \text{or } L^{W_1}\left(e_{ij}^{(1)}\right)\neq L^{W_1}\left(e_{ij}^{(2)}\right)\right), \Rightarrow l^*\left(e_{ij}^{(1)}\right)=\phi$,

then the distance under the correspondence l^* is minimum, namely becomes the distance between the attributed relational graphs $I^{(1)}$ and $I^{(2)}$, where it is called as optimal correspondence between $I^{(1)}$ and $I^{(2)}$.

The optimal correspondence between attributed relational graphs means that the corresponding nodes have the same node's labels (or haven't corresponding nodes) and the corresponding edges have the same edge's labels (or haven't corresponding edges).

It should be noted that the distance of with the correspondence $l: I^{(1)} \rightarrow I^{(2)}$ between $I^{(1)}$ and $I^{(2)}$ is not equal to the distance of with the inverse correspondence $l^{-1}: I^{(2)} \rightarrow I^{(1)}$, i.e. $d_t(I^{(1)}, I^{(2)}|l) \neq d_t(I^{(2)}, I^{(1)}|l^{-1})$ (where $t \in \{V,E\}$). It is due to the existence of non-corresponding nodes or edges. Therefore, virtual nodes and edges are added to the attributed relational graphs instead of the non-corresponding nodes and edges, so as to satisfy the symmetry of distance. Here, the values of attributes of the virtual nodes and edges are zero. Moreover, the distance between sets of edge's attributes is normalized to meet $d_E(I^{(1)}, I^{(2)}|l) \in [0,1]$.

Figure 2 shows the image description model and distance measure based on attributed relational graph. The distance between sets of node's attributes reflects the mismatching degree for semantic object information between images and it is relative to the COV-based distance proposed by Vogel et al. [32]. The distance between sets of edge's attributes reflects the mismatching degree for spatial structure information between images. The semantic distance reflects the degree of matching between two ARGs for semantic objects contained in images and spatial relations between objects. The proposed image description model and distance measure method are based on human perception mechanism for image understanding.

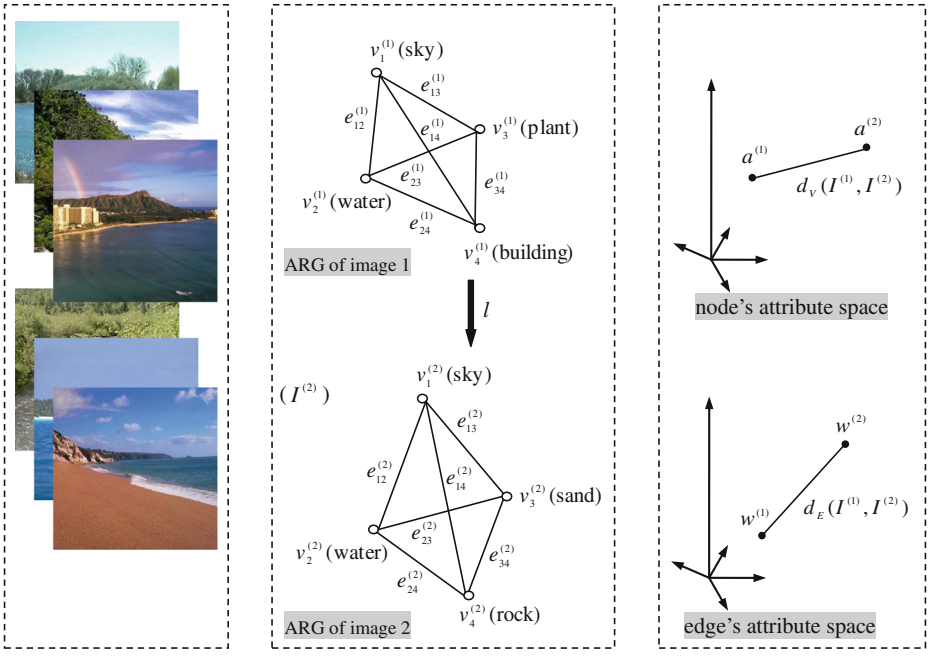


Fig. 2 Image description and distance measure based on attributed relational graph. **a** Original images. **b** Attributed relational Graphs. **c** Distances in two attribute spaces

3 Image classification with Bayesian network

In training stage, firstly, the training images are described as attributed relational graphs, namely the set of semantic objects and the set of relations between objects are extracted and their attributes are computed. Then two kinds of SVM classifiers are trained with distance measure method based on attributed relational graphs: one is the object SVM classifiers defined in node's attribute feature space; another is the structure SVM classifiers defined in edge's attribute feature space. The object SVM classifiers are based on information of semantic objects in an image and are constructed with the distance between sets of node's attributes, while the structure SVM classifiers are based on information of spatial structure of an image and are constructed with the distance between sets of edge's attributes. Afterwards, by using the trained object SVM classifiers and structure SVM classifiers, the joint conditional probability distributions of Bayesian network are computed, which include the object conditional probability and structure conditional probability. In test stage, the test images are described as attributed relational graphs. Then, using the joint conditional probability distributions, we want to find the image category with maximum a posterior probability.

3.1 Construction of the structure of Bayesian network

Bayesian network which is used to encode the dependence relations between random variables is a directed acyclic graph (DAG). Here, the nodes denote the random variable and the directed edges denote the relations between parents and children. Formally, Bayesian network is defined as $B=(G,\Theta)$,where G is the network structure and Θ is the conditional probability matrix. Given a set $U=\{X_1,X_2,\ldots,X_n\}$ of random variables, where the values of each variable are expressed as $Val(X_i)$, then the Bayesian network defined in U encodes the dependence relations between random variables $X_1,X_2,\ldots,X_n$. If there is a directed edge from node X_i to node X_j, then X_i is called as a parent of X_j. Each node has a conditional probability distribution $p(X_i|parents(X_i))$, which quantitatively presents the influence of parent nodes. In Bayesian network, the conditional probability between nodes is acquired by learning methods and posterior probability for new evidence is calculated with prior and conditional probability distributions.

In order to construct the structure of Bayesian network for a specified field, the variables of network and their values are generally decided by domain experts, then the dependence relations between variables are determined using manual or learning methods. When it is difficult to determine the network structure manually, the learning methods are used to automatically construct the network structure from training data set. But, the learning methods for Bayesian network structure are still in the initial stage. However, in the case of the clear causal relations, the structure of Bayesian network can be determined with domain knowledge of the relations between different entities.

The structure of Bayesian network for image classification is constructed based on analyzing of relations between elements in an image scene. An image scene is consisted of a collection of objects and relations between objects. First of all, we define a root node to represent the image categories. Then, we define the object nodes and the relation nodes, and decide the causal relations among nodes. The objects and relations between them are affected by image category, so the root node is a parent of all object nodes and relation nodes. We assume that the occurence of an object does not affect the occurence of another object in an image, so the object nodes are independent on each other. Similarly, the relation nodes are also independent on each other. Moreover, the relations between objects should be depended on two corresponding objects, so each pair of object nodes is a parent of relation node between them.

Figure 3 shows the structure of Bayesian network:root node C represents the image categories; nodes $\{A_1,A_2,\ldots,A_m\}$ ($A_i \in A$) and $\{W_{12},W_{13},\ldots,W_{m-1\ m}\}$ denote the objects in an image and the spatial relations between objects, respectively; root node C is a parent of object node A_i; root node C and object nodes A_i, A_j are a parent of relation node W_{ij}. Here, the root node takes on values from a set of image categories $\{C_1,C_2,\ldots,C_N\}$; the object nodes and the relation nodes take on values from a set of object attributes a and a set of relation attributes w, respectively.

3.2 Learning the conditional probabilities

In general case, there are two kinds of method to obtain the conditional probability matrix of Bayesian network, i.e. expert knowledge based method and learning based method. The learning based method generally estimates the conditional probabilities from the training data set by calculating the frequencies of random events. Namely, for image classification, the conditional probabilities $p(A_i|C_n)$ are estimated by calculating the frequencies of semantic objects A_i in image categories C_n:

$$\vartheta_{ni}^{A} = p\left(A_i \middle| C_n\right) = \frac{\sum_{I \in C_n} a_i(I)}{\Sigma_{C_n}} \tag{13}$$

where, Σ_{C_n} is the number of images which belong to the category C_n from a training data set, and it satisfies the condition $\sum_i p(A_i|C_n) = 1$. Let the number of image categories be N and the number of semantic objects be M, then the conditional probability matrix of objects is a $N \times M$ form matrix: $\Theta^A = [\vartheta_{ni}^A]_{N \times M}$.

The conditional probabilities $p(W_{k,ij}|C_n,A_i,A_j)$ are estimated by calculating the frequencies of spatial relations $W_{k,ij}$ between objects A_i and A_j in image categories C_n:

$$\vartheta_{nk}^{W} = p\left(W_{k,ij} \middle| C_n, A_i, A_j\right) = \frac{\sum_{I \in C_n} w_{k,ij}(I)}{\Sigma_{C_n}} \tag{14}$$

The number of possible combinations of (C_n, A_i, A_j) is $R = N \cdot C_M^2$, and the conditional probability matrix of spatial relations is a $R \times 3$ form matrix: $\Theta^W = [\vartheta_{nr}^W]_{R \times 3}$.

The method of estimating the conditional probability by calculating the frequencies needs a lot of ground truth data. In sparse training data case, the accuracy of calculated conditional probabilities may be insufficient. Therefore, we use the kernel method to solve this problem.

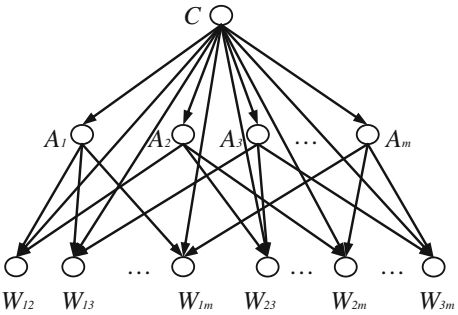


Fig. 3 Structure of the Bayesian network

Kernel method is based on introducing the kernel function instead of the inner product function in the high dimensional space. In fact, the kernel function reflects the similarity measure between image models for image classification. Therefore, by training the SVM classifiers, we obtain the optimal classification samples which distinguish the image categories, i.e. support vectors. Using the distance measure between images based on attributed relational graph, we obtain the joint conditional probability with kernel method. In order to train the SVM classifiers, we don't use the original feature model of an image, instead, the distances between proposed image models are calculated.

Given a set of training data $\{(I^{(i)}, C_n) | I^{(i)}=(V^{(i)}, E^{(i)}, a^{(i)}, w^{(i)}),\ C_n \in C\}_{i=1}^{\Sigma}$, we divide the semantic features of an image into a feature of semantic objects $\{((V^{(i)}, a^{(i)}), C_n)\}_{i=1}^{\Sigma}$ and a feature of saptial relations $\{((E^{(i)}, w^{(i)}), C_n)\}_{i=1}^{\Sigma}$ between objects. Training the SVM classifiers based on semantic distance between images mentioned above, we can obtain the separation hyperplanes, which separate the images into a certain category C_n against the remaining categories. Here, we use a non-linear Gaussian RBF kernel based on its performance in other pattern recognition applications [27]. We train the two kinds of SVM classifiers and their kernel functions are given as follows:

$$K_t\left(I^{(i)}, I^{(j)}\right) = \exp\left\{-\frac{d_t\left(I^{(i)}, I^{(j)}\right)}{2\sigma_t^2}\right\}, t \in \{V, E\} \tag{15}$$

where, σ_t is the kernel radius. Let $d_V(C_n|a)$ and $d_E(C_n|w)$ denote the distances between the image $I=(V,E,a,w)$ and the hyperplane of category C_n in two kinds of feature space, they can be transformed to probability spaces using sigmoid function similar with [27]:

$$p\left(C_n \middle| a\right) = 1/\left(1 + \exp\left(-\gamma \cdot d_V\left(C_n \middle| a\right)\right)\right) \tag{16}$$

$$p\left(C_n \middle| w\right) = 1/\left(1 + \exp\left(-\gamma \cdot d_E\left(C_n \middle| w\right)\right)\right) \tag{17}$$

where, γ is the slope parameter of the sigmoid function. By using Bayesian rule, the joint conditional probability $p(A_1,A_2,\ldots,A_M | C_n)$ is modified with the posterior probability obtained by SVM classifier:

$$p\left(A_1, A_2, \ldots, A_M \middle| C_n\right) = p\left(a \middle| C_n\right) = \frac{p(a)}{p(C_n)} p\left(C_n \middle| a\right) \tag{18}$$

where, $p(a)$ and $p(C_n)$ are the prior distributions of object attributes and scenes. Here we model the prior probabilities with simply a flat prior, then we have the joint conditional probability of object nodes as folows:

$$p\left(A_1, A_2, \ldots, A_M \middle| C_n\right) \propto p\left(C_n \middle| a\right) = 1/\left(1 + \exp\left(-\gamma \cdot d_V\left(C_n \middle| a\right)\right)\right) \tag{19}$$

Similarly, we have the joint conditional probability of spatial relation nodes as follows:

$$p\left(W_{12}, W_{13}, \ldots, W_{M-1M} \middle| C_n, A\right) \propto p\left(C_n \middle| w\right) = 1/\left(1 + \exp\left(-\gamma \cdot d_E\left(C_n \middle| w\right)\right)\right) \tag{20}$$

In proposed method, there is no need to calculate the individual conditional probabilities, instead, it directly obtains the joint conditional probabilitiy, which is enough to infer the image categiry. Here, we use the optimal classification samples (i.e. support vectors) to learn the conditional probability, which reflects that the object attributes and spatial structure attributes

of an image belong to a certain category. Therefore, it ensures the accuracy of conditional probability distribution with sparse training samples.

3.3 Image classification

The test image is segmented into different image regions, and each region is annotated as one of the nine semantic objects, so we obtain the set $A=\{A_1,A_2,\ldots,A_M\}$ of objects of the test image. Using the Eq. (2), we obtain the set a of semantic object attributes of the test image. We calculate the fuzzy membership degrees of all spatial relations between objects of the tast image, so we obtain the set $W=\{W_{12},W_{13},\ldots,W_{M-1M}\}$ of spatial relations and the set w of their attributes. Then, we carry out the classification of the test image with Bayesian network.

Using Bayesian network, we want to find the image category with maximum a posterior probability, given the input evidence of the test image. Let $\{A, W\}$ be the input evidence of Bayesian network, then semantic category of the test image is as follows:

$$C^* = arg\max_{C_n} p\left(C_n \middle| A, W\right) \tag{21}$$

By exploiting Bayesian rule and Markov independence condition, we have

$$\begin{aligned} C^* &= arg\max_{C_n} p(C_n)p\left(A_1, A_2, \ldots, A_M \middle| C_n\right)p\left(W_{12}, W_{13}, \ldots, W_{M-1M} \middle| C_n, A\right) \\ &= arg\max_{C_n} p\left(C_n\right) \prod_{i=1}^{M} p\left(A_i \middle| C_n\right) \prod_{\substack{i,j=1 \\ i\neq j, i<j}}^{M} \prod_{k=1}^{3} p\left(W_{k,ij} \middle| C_n, A_i, A_j\right) \end{aligned} \tag{22}$$

The corresponding derivation process is provided in the Appendix.

The conditional probability $p(A \mid C_n)$ and $p(W \mid C_n)$ are obtained by the traditional method (Eqs. (13) and (14)) or the proposed SVM-based method (Eqs. (19) and (20)). The terms of the first row in Eq. (22) mean the joint conditional probability obtained by proposed method, while the terms of the second row in Eq. (22) mean the individual conditional probability obtained by traditional method.

Finally, the category of the test image inferred by Bayesian network is expressed as follows:

$$\begin{aligned} C^* &= arg\max_{C_n} p\left(C_n\right) p\left(Val(A_1) = a_1, Val(A_2) = a_2, \ldots, Val(A_M) = a_M \middle| C_n\right) \\ &\quad p\left(Val(W_{12}) = w_{12}, Val(W_{13}) = w_{13}, \ldots, Val(W_{M-1M}) = w_{M-1M} \middle| C_n, A\right) \\ &= arg\max_{C_n} p(C_n) \prod_{i=1}^{M} p\left(Val(A_i) = a_i \middle| C_n\right) \prod_{\substack{i,j=1 \\ i\neq j, i<j}}^{M} \prod_{k=1}^{3} p\left(Val\left(W_{k,ij}\right) = w_{k,ij} \middle| C_n, A_i, A_j\right) \end{aligned} \tag{23}$$

4 Experimental results

4.1 Experimental setup

In order to verify the effectiveness of image classification method proposed in this work, we carried out the related experiments with two fields: manually annotated regions and automatically classified regions. We used four benchmark image databases as follows:

1) image dataset provided by Vogel and Schiele [32] (VS dataset), which contains 700 images of 720*480 pixels resolution and is classified 6 natural scene categories as follows: coasts (142 images), river/lakes (111 images), forests (103 images), plains (131 images), mountains (179 images) and sky/clouds (34 images);
2) image dataset provided by Fei-Fei and Perona [14] (FP dataset) containing a part of Corel image dataset, which contains 3759 images in 13 categories of natural scenes: highway (260 images), inside of cities (308 images), tall buildings (356 images), streets (292 images), suburb residence (241 images), forest (328 images), coast (360 images), mountain (374 images), open country (410 images), bedroom (174 images), kitchen (151 images), livingroom (289 images) and office (216 images), and the resolution of each image is approximately 250*300 pixels;
3) the image dataset *spatial_envelope_256× 256_static_8outdoorcategorie* provided by LabelMe (http://labelme.csail.mit.edu/), which contains 1418 images of 256*256 pixels resolution and is classified into 5 image categories: beach (236 images), forest (325 images), mountain (374 images), field (364 images) and river/lake (119 images);
4) Caltech-101 (http://www.vision.caltech.edu/) image dataset, in which the images of objects classified into 101 categories and each category contains about 40 to 800 images of roughly 300×200 pixels resolution;

While the image segmentation and region annotaion are no major topics of this work, we utilized the advanced stability-based clustering method for image segmentation which proposed in Rabinovich et al. [26]. Then we used the discriminative concept classifiers [32] which classify the local image regions into the semantic objects or meterials. Where, the visual feature vectors extracted from the image regions are the concatenations of a 54-bin linear HSV color histogram, an 8-bin edge direction histogram and the 24 features of the gray-level co-occurrence matrix: contrast, energy, entropy, homogeneity, inverse difference moment and correlation for the displacements $\overrightarrow{1,0}$, $\overrightarrow{1,1}$, $\overrightarrow{0,1}$ and $\overrightarrow{-1,1}$ [11].

As mentioned in the previous section, we used the SVM classifiers with a non-linear Gaussian RBF kernel to obtain the joint conditional probability. The values for kernel radiuses σ_V, σ_E and cost parameteres [27] c_V, c_E of SVM were shown in Table 1, which have been slected empirically whith 10-fold cross validation to maximize the average performance of classification, and the other parameter values used in this experiment are also were shown in Table 1.

In each image category, we select η% of images that are regarded as a set of training images, and the rest (100 - η) % of images are regarded as a set of test images. In order to make a fair comparison, all the experiments are carried out by a 10-fold cross-validation process on different training and test sets. The reported results are the average values of them. The classification performance is evaluated with the classification accuracy and the confusion matrix.

Table 1 The parameter values used in these experiments

Parameter	α_1	β_1	α_2	β_2	γ	σ_V	σ_E	c_V	c_E
Value	20	0.25	10	0.6	1.0	2	1.5	10	10

4.2 Classification results

4.2.1 Comparison of proposed model with state-of-the-art baselines

In order to verify the effectiveness of the proposed image description model and distance measure method which are based on attributed relational graph, we compared the image classification results with several state-of-the-art baselines, namely Vogel and Schiele [32], Boutell et al. [6], Cheng and Wang [11], Liu Shuoyan et al.[21], Jin Biao et al. [19] and Oilivier Duchenne et al. [13] for VS, FP and Caltech-101 image datasets. Table 2 shows the comparison of average classification accuracies where η=50 % with the baselines.

In the COV-based model by Vogel and Schiele [32], an image was divided into 10 * 10 blocks, and on each block a feature vector combined the several visual features was extracted. Using the SVM classifiers, each block is classified into one of the predefined nine semantic objects. Then, the 9-dimensional concept occurrence vector was used as input to SVM classifiers, which classify an image into individual image categories. In the Factor-Graph model by Boutell et al. [6], an image was firstly segmented and labeled by several objects. Then they have modeled the pairwise relationships between regions and estimated the scene probabilities using loopy belief propagation on a factor graph. In the CBN model by Cheng and Wang [11], the hybrid streams of object occurrence information and spatial arrangement information of image were piped into the CBN-based inference engine. They have used the spatial relations between objects which were computed through the key objects. In the R-CRF (Region-Conditional Random Fields) model by Liu Shuoyan et al. [21], they appended the region contextual constraint into traditional BoW framework. They introduced R-CRF model, where the potential function was built under the region contextual constraint, to learn each visual word depending on the rest of the visual words in the same region. In the SR-pLSA model by Jin Biao et al. [19], they used a histogram to describe the spatial relations and classified them into spatial relation labels (left, right, above, below, near, far, inside, outside). Then they extended the probabilistic latent semantic analysis (pLSA) by taking into account the spatial relationships between topics (SR-pLSA), and employed the SVM of which input is the SR-pLSA to classify the images. In the Graph-Matching Kernel model by Oilivier Duchenne et al. [13], they defined the matching problem between two images as an optimization of energy in multi-label Markov random fields (MRF) which are defined on the

Table 2 Comparison of the average classification accuracy with several state-of-the-art baselines (%)

Datasets	VS dataset		FP dataset	Caltech-101
	Manual annotated regions	Automatically classified regions		
COV-based model by Vogel and Schiele [32] (2007)	86.4	74.1		
Factor-Graph model by Boutell et al. [6] (2007)	82.3	68.4		
CBN model by Cheng and Wang [11] (2010)	88.7	75.9		
R-CRF model by Liu Shuoyan et al. [21] (2011)			74.5	
SR-pLSA model by Jin Biao et al. [19] (2012)	88.9		86.5	
Graph-Matching Kernel by Oilivier Duchenne et al. [13] (2011)				80.3
ARG-based model proposed in this work	**90.0**	78.7	**86.8**	**82.1**

Data in bold emphasis represent the best values of the classification accuracy

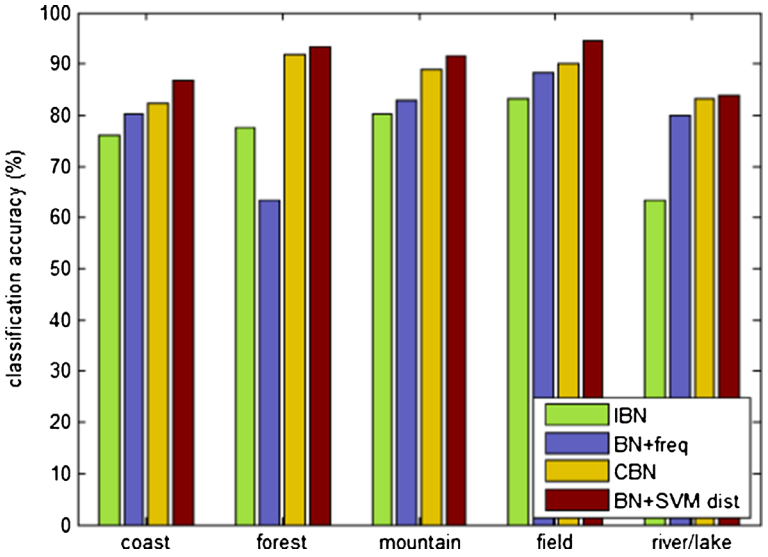


Fig. 4 Comparison of the classification accuracy of images using different methods with Bayesian network—based on manual annotated regions

corresponding graphs. Then they considered the value of the optimized MRF associated with two images as a kernel, performed image classification by SVM classifier.

Using the image description model based on attributed relational graph (ARG), we obtained the object attributes and the spatial structure attributes of an image. Then, according to the distance measure method between the attributed relational graphs, we carried out the image classification with SVM and Bayesian classifiers.

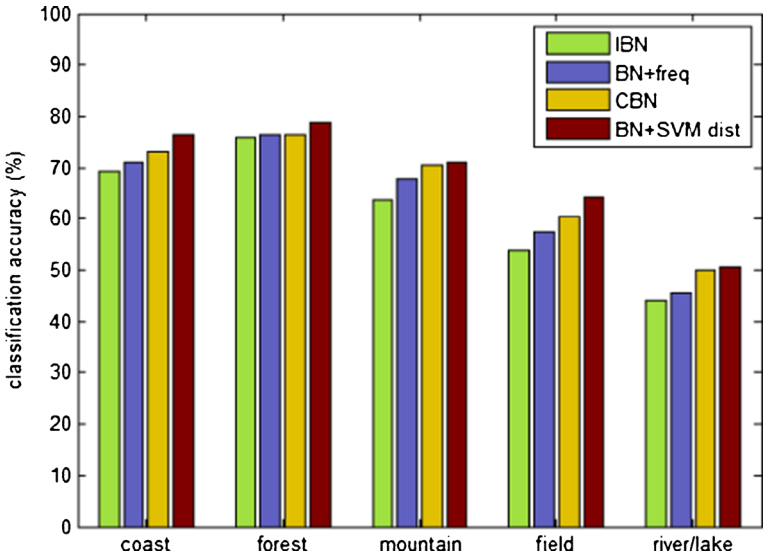


Fig. 5 Comparison of the classification accuracy of images using different methods with Bayesian network—based on classified regions

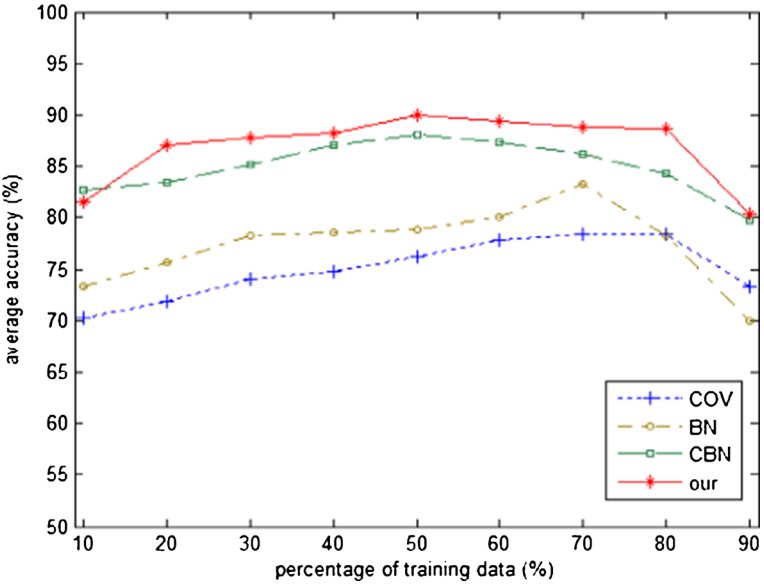


Fig. 6 Change curve of the average classification accuracy of images with the change of the η

As shown in Table 2, the proposed ARG-based method nearly outperforms the state-of-the-art baselines, except the SR-pLSA model for automatically classified regions on VS dataset. Specifically, the average classification accuracy increased 1 %~8 % for manual annotated regions and 3 %~10 % for automatically classified regions of VS dataset respectively. This is related that the proposed ARG-based method not only uses the semantic object information, but also uses the fully-connected spatial structure information, while the COV-based method only considers the semantic object information of an image, and Factor-Graph model and CBN model merely consider the spatial relations through adjacent pair-wise regions or key object regions, which not fully reflect the spatial arrangement of an image. Compared with the SR-pLSA model, our classification accuracy was decreased a little for automatically classified regions. This is caused by imperfect result of image segmentation and region annotation, yet the image classification result strongly depends on the performance of image segmentation and region annotation. Making full use of the spatial relations between objects, the proposed ARG-based method outperforms the R-CRF model based method, SR-pLSA method and the Graph-Matching Kernel based method for FP and Caltech-101 image datasets. Compared with the Graph-Matching Kernel method by Oilivier et al. [13], the average classification accuracy increased about 2 % for Caltech-101 dataset. The Graph-

Table 3 Confusion matrix of the image classification using COV-based method

	Coast	Forest	Mountain	Field	River/lake
Coast	**75.1**	0.0	3.6	2.0	19.3
Forest	0.0	**75.0**	4.1	15.3	5.6
Mountain	0.0	12.1	**82.8**	3.4	1.7
Field	1.8	7.8	1.4	**86.4**	2.6
River/lake	13.7	7.9	6.6	10.1	**61.7**

Data in bold emphasis represent the maximum value in each row of confusion matrix

Table 4 Confusion matrix of the image classification using BN-based method

	Coast	Forest	Mountain	Field	River/lake
Coast	**80.3**	0.0	0.0	1.6	18.1
Forest	0.0	**63.3**	8.3	11.7	16.7
Mountain	1.7	1.7	**82.8**	6.9	6.9
Field	1.8	3.2	0.0	**88.2**	6.8
River/lake	13.3	1.7	0.0	5.0	**80.0**

Data in bold emphasis represent the maximum value in each row of confusion matrix

Matching Kernel approach [13] was based on a graphical description model of image content as our approach: an image is represented by graph of which nodes and edges correspond to the regions and their spatial relations. The differences between the method of Oilivier et al. and our method are mainly that we used not regular image patch but segmented semantic objects and investigate the spatial arrangements in an image more sufficiently by defining fuzzy contextual relations, and we employed SVM and Bayesian network with proposed ARG-based semantic distance, while Oilivier et al. employed SVM with Graph-Matching distance for image classification.

4.2.2 Comparison between performances by Bayesian network models

In order to verify the effectiveness of the porposed probability calculation method of Bayesian network, we compare the experimental results btween four image classification methods using Bayesian network for LabelMe dataset, namely, (1) IBN (Indipendent Bayesian network): this does not consider the spatial relations between objects, where the network structure contains only the object nodes, and it is similar to the indoor/outdoor classification method proposed by Serrano et al.[27]; (2) BN + freq: where the conditional probabilities are estimated with calculating the frequencies by Eqs. (13) and (14); CBN: it is the image classification method based on contextual Bayesian network proposed by Cheng et al. [11]; (4) BN + SVM dist: it is the SVM-based method proposed in this work, where the joint conditional probabilities are calculated by Eqs. (19) and (20).

Figure 4 shows the classification results based on manual annotated regions, and Fig. 5 shows the classification results based on automatically classified regions, where η=50 %. The automatic region annotation was obtained by energy based method with the contextual model based on conditional random field [16]. The proposed SVM-based method with ARG distance outperforms other Bayesian network models. Since the object recognition accuracy have been substantially low (less than 50 %), the classification results based on automatic region annotation are lower than manual annotation case. Therefore, we use the manual annotated regions to analyze the relations between a classification accuracy and η.

Table 5 Confusion matrix of the image classification using CBN-based method

	Coast	Forest	Mountain	Field	River/lake
Coast	**82.4**	0.0	0.0	2.0	15.6
Forest	0.0	**91.7**	0.0	4.3	4.0
Mountain	0.0	3.4	**88.8**	3.9	3.9
Field	0.0	3.3	0.0	**90.0**	6.7
River/lake	13.3	1.7	0.0	1.7	**83.3**

Data in bold emphasis represent the maximum value in each row of confusion matrix

Table 6 Confusion matrix of the image classification using ARG-based method proposed in this work

	Coast	Forest	Mountain	Field	River/lake
Coast	**86.9**	0.0	0.0	0.0	13.1
Forest	0.0	**93.3**	1.7	1.7	3.3
Mountain	0.0	5.2	**91.4**	1.7	1.7
Field	1.8	3.6	0.0	**94.6**	0.0
River/lake	6.3	5.0	1.7	3.3	**83.7**

Data in bold emphasis represent the maximum value in each row of confusion matrix

Figure 6 shows the curves of average accuracy of four classification methods based on Bayesian network. The experimental results show that the accuracy of BN + freq method was improved with the increasing of the number of training data except $\eta<30$ % . But the accuracy of the BN + SVM dist was significantly higher than the other three methods in $20\% \leq \eta \leq 80\%$. This illustrates that the SVM-based method with attributed relational graph is suitable and effective to estimate the conditional probability for sparse training samples. However, in the case of too little number of training data, the learning results by SVM are not so well. And, in the case of too much number of training data, the number of test data is too little, so it may lead to the increase of the effect of each test result to the overall classification performance and the average accuracy may be decreased.

The confusion matrices of four kinds of semantic-based image classification methods are shown in Tables 3, 4, 5 and 6, where $\eta=50$ %. The diagonal elements of the matrices represent a classification accuracy. Comparing with COV-based method, BN-based method and CBN-based method, the average classification accuracy of the method proposed in this work was improved to 13.06 %, 11.06 % and 2.74 %, respectively.

Through the comparison analysis of classification experiments, we have verified the suitability and effectiveness of the Bayesian network based semantic image classification method with attributed relational graph proposed in this work. The proposed method makes full use of the object information and spatial relational information in an image. In addition, it ensures the accuracy of joint conditional probability with sparse training samples. Therefore, it improves the performance of image classification.

5 Conclusions

It is an important means to improve the accuracy of semantic-based image classification, that make full use of a semantic object information and a spatial structure information in an image. In this paper, we proposed an image description model based on attributed relational graph and a semantic distance measure method between images, thus we have made full use of information of the semantic objects and spatial relations between objects in an image. In addition, to ensure the accuracy of probability distribution with sparse training samples, we proposed a method which is based on SVM and proposed distance measure. Therefore we have improved the accuracy of image classification. Experimental results showed the validity and the suitability of the proposed method.

However, the image classification results depend on the performance of image segmentation and region annotation. An image segmentation and region annotation with high semantic level have been difficult so far. In futher, we will research the semantic-baseed image

classification which is closely combined with the image segmentation and region annotation, so as to achieve better classification results.

Acknowledgments This work was partly supported by the 973 Program (2013CB329504), NSF of China (No. 61272261), NSF of Zhejiang (Y1110152), and STD of Zhejiang (2012C21002).

Appendix: Derivation of image classification formula with Bayesian network

$$
\begin{aligned}
C^* &= arg\max_{C_n} p\left(C_n \middle| A, W\right) \\
&= arg\max_{C_n} \frac{p(C_n)p\left(A, W \middle| C_n\right)}{p(A, W)} = arg\max_{C_n} \frac{p(C_n)p\left(A \middle| C_n\right)p\left(W \middle| C_n, A\right)}{p(A, W)} \\
&= arg\max_{C_n} p(C_n)p\left(A \middle| C_n\right)p\left(W \middle| C_n, A\right) \\
&= arg\max_{C_n} p(C_n)p\left(A_1, A_2, \ldots, A_M \middle| C_n\right)p\left(W_{12}, W_{13}, \ldots, W_{M-1M} \middle| C_n, A\right) \\
&= arg\max_{C_n} p(C_n)\prod_{i=1}^{M} p\left(A_i \middle| C_n\right) \prod_{\substack{i,j=1 \\ i\neq j, i<j}}^{M} p\left(W_{ij} \middle| C_n, A_i, A_j\right) \\
&= arg\max_{C_n} p(C_n)\prod_{i=1}^{M} p\left(A_i \middle| C_n\right) \prod_{\substack{i,j=1 \\ i\neq j, i<j}}^{M} p\left(W_{1,ij}, W_{2,ij}, W_{3,ij} \middle| C_n, A_i, A_j\right) \\
&= arg\max_{C_n} p(C_n)\prod_{i=1}^{M} p\left(A_i \middle| C_n\right) \prod_{\substack{i,j=1 \\ i\neq j, i<j}}^{M} \prod_{k=1}^{3} p\left(W_{k,ij} \middle| C_n, A_i, A_j\right)
\end{aligned}
$$

References

1. Aksoy S, Koperski K, Tusk C et al (2005) Learning Bayesian classifiers for scene classification with a visual grammar [J]. IEEE Trans Geosci Remote Sens 43(3):581–589. doi:10.1109/TGRS.2004.839547
2. Aksoy S, Tusk C, Koperski K, et al. (2003) Scene modeling and image mining with a visual grammar [J]. Frontiers of Remote Sensing Information Processing, 35–62. doi:10.1142/9789812796752_0003
3. Athanasiadis T, Mylonas P, Avrithis Y et al (2007) Semantic image segmentation and object labeling [J]. Circ Syst Video Technol IEEE Trans 17(3):298–312. doi:10.1109/TCSVT.2007.890636
4. Berretti S, Del Bimbo A, Vicario E (2001) Efficient matching and indexing of graph models in content-based retrieval [J]. IEEE Trans Patt Anal Mach Intell 23(10):1089–1105. doi:10.1109/34.954600
5. Bosch A, Zisserman A, Muoz X (2008) Scene classification using a hybrid generative/discriminative approach [J]. Patt Anal Mach Intell IEEE Trans 30(4):712–727. doi:10.1109/TPAMI.2007.70716
6. Boutell MR, Luo J, Brown CM (2007) Scene parsing using region-based generative models [J]. IEEE Trans Multimedia 9(1):136–146. doi:10.1109/tmm.2006.886372
7. Bruzzone L, Persello C (2009) A novel context-sensitive semisupervised SVM classifier robust to mislabeled training samples [J]. IEEE Trans Geosci Remote Sens 47(7):2142–2154. doi:10.1109/tgrs.2008.2011983
8. Caputo B, Jie L (2009) A performance evaluation of exact and approximate match kernels for object recognition [J]. Electron Lett Comp Vision Image Anal 8(3):15–26

9. Carbonetto P, de Freitas N, Barnard K (2004) A statistical model for general contextual object recognition[C]. Computer Vision - ECCV 2004. Lecture Notes in Computer Science Volume 3021: 350–362. doi:10.1007/978-3-540-24670-1_27
10. Chang E, Goh K, Sychay G et al (2003) CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines [J]. Circ Syst Video Technol IEEE Trans 13(1):26–38. doi:10.1109/TCSVT.2002.808079
11. Cheng H, Wang R (2010) Semantic modeling of natural scenes based on contextual Bayesian networks [J]. Patt Recog 43(12):4042–4054. doi:10.1016/j.patcog.2010.06.004
12. Choi MJ, Torralba A, Willsky AS (2012) A tree-based context model for object recognition [J]. Patt Anal Mach Intell IEEE Trans 34(2):240–252. doi:10.1109/TPAMI.2011.119
13. Duchenne O, Joulin A, Ponce J (2011) A graph-matching kernel for object categorization[C]//Proc IEEE Int Conf Comp Vis (ICCV), 1792–1799. doi:10.1109/ICCV.2011.6126445
14. Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories [C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2: 524–531. doi:10.1109/CVPR.2005.16
15. Galleguillos C, Belongie S (2010) Context based object categorization: a critical survey [J]. Comp Vision Image Underst 114(6):712–722. doi:10.1016/j.cviu.2010.02.004
16. Galleguillos C, Rabinovich A, Belongie S (2008) Object categorization using co-occurrence, location and appearance [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–8. doi:10.1109/CVPR.2008.4587799
17. Harchaoui Z, Bach F (2007) Image classification with segmentation graph kernels[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), 1–8. doi:10.1109/CVPR.2007.383049
18. Huang Y, Huang K, Wang C, et al. (2011) Exploring relations of visual codes for image classification[C]// Computer Vision and Pattern Recognition (CVPR). IEEE Conference on. IEEE, 1649–1656. doi:10.1109/CVPR.2011.5995655
19. Jin B, Hu W, Wang H (2012) Image classification based on pLSA fusing spatial relationships between Topics [J]. Signal Process Lett IEEE 19(3):151–154. doi:10.1109/LSP.2012.2184091
20. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C]//Computer Vision and Pattern Recognition. IEEE Comp Soc Conf IEEE 2:2169–2178. doi:10.1109/CVPR.2006.68
21. Liu S, Xu D, Feng S (2011) Region contextual visual words for scene categorization [J]. Expert Syst Appl 38(9):11591–11597. doi:10.1016/j.eswa.2011.03.037
22. Luo J, Savakis AE, Singhal A (2005) A Bayesian network-based framework for semantic image understanding [J]. Patt Recog 38(6):919–934. doi:10.1016/j.patcog.2004.11.001
23. Papadopoulos GT, Saathoff C, Escalante HJ et al (2011) A comparative study of object-level spatial context techniques for semantic image analysis [J]. Comp Vision Image Underst 115(9):1288–1307. doi:10.1016/j.cviu.2011.05.005
24. Park BG, Lee KM, Lee SU et al (2003) Recognition of partially occluded objects using probabilistic ARG (attributed relational graph)-based matching [J]. Comp Vision Image Underst 90(3):217–241. doi:10.1016/S1077-3142(03)00049-3
25. Qi GJ, Hua XS, Rui Y et al (2010) Image classification with kernelized spatial-context [J]. IEEE Trans Multimedia 12(4):278–287. doi:10.1109/TMM.2010.2046270
26. Rabinovich A, Vedaldi A, Galleguillos C, et al. (2007) Objects in context[C]//In Proc. of 11th International Conference on Computer Vision, 1–8. doi:10.1109/ICCV.2007.4408986
27. Serrano N, Savakis AE, Luo J (2004) Improved scene classification using efficient low-level features and semantic cues [J]. Patt Recog 37(9):1773–1784. doi:10.1016/j.patcog.2004.03.003
28. Shen J, Shepherd J, Ngu AHH (2005) Semantic-sensitive classification for large image libraries[C]// Multimedia Modelling Conference, 2005. MMM 2005. Proceedings of the 11th International. IEEE, 340–345. doi:10.1109/MMMC.2005.66
29. Singhal A, Luo J, Zhu W. Probabilistic spatial context models for scene content understanding [C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, 1: 235–241. doi:10.1109/CVPR.2003.1211359
30. Su Y, Jurie F (2012) Improving image classification using semantic attributes [J]. Int J Comp Vision 100(1):59–77. doi:10.1007/s11263-012-0529-4
31. Vailaya A, Figueiredo MAT, Jain AK et al (2001) Image classification for content-based indexing [J]. Image Proc IEEE Trans 10(1):117–130. doi:10.1109/83.892448
32. Vogel J, Schiele B (2007) Semantic modeling of natural scenes for content-based image retrieval [J]. Int J Comp Vision 72(2):133–157. doi:10.1007/s11263-006-8614-1

33. Yang J, Yu K, Gong Y, et al. (2009) Linear spatial pyramid matching using sparse coding for image classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 1794–1801. doi:10.1109/CVPR.2009.5206757
34. Yang L, Zheng N, Yang J (2011) A unified context assessing model for object categorization [J]. Comp Vision Image Underst 115(3):310–322. doi:10.1016/j.cviu.2010.10.011

Chang-yong Ri received the B.S. and M.S. degrees in information and computer science from Kim Il Song University, Pyongyang, D.P.R. of Korea. He is currently a Ph.D. degree candidate of the School of Computer Science and Technology, Zhejiang University, Hangzhou, China. His research interests include semantic processing in image, content based retrieval and pattern recognition.

Min Yao is a Professor and Ph.D. supervisor of the School of Computer Science and Technology, Zhejiang University, Hangzhou, China. His research interests include computational intelligence, image processing and pattern recognition.