Image Classification With Kernelized Spatial-Context

Guo-Jun Qi, Xian-Sheng Hua, *Member, IEEE*, Yong Rui, *Fellow, IEEE*, Jinhui Tang, *Member, IEEE*, and Hong-Jiang Zhang, *Fellow, IEEE*

***Abstract*—The goal of image classification is to classify a collection of unlabeled images into a set of semantic classes. Many methods have been proposed to approach this goal by leveraging visual appearances of local patches in images. However, the spatial context between these local patches also provides significant information to improve the classification accuracy. Traditional spatial contextual models, such as two-dimensional hidden Markov model, attempt to construct one common model for each image category to depict the spatial structures of the images in this class. However due to large intra-class variances in an image category, one single model has difficulties in representing various spatial contexts in different images. In contrast, we propose to construct a prototype set of spatial contextual models by leveraging the kernel methods rather than only one model. Such an algorithm combines the advantages of rich representation ability of spatial contextual models as well as the powerful classification ability of kernel method. In particular, we propose a new distance measure between different spatial contextual models by integrating joint appearance-spatial image features. Such a distance measure can be efficiently computed in a recursive formulation that scales well to image size. Extensive experiments demonstrate that the proposed approach significantly outperforms the state-of-the-art approaches.**

***Index Terms*—2-D hidden Markov model, image classification, kernel method, spatial context.**

I. Introduction

IMAGE categorization has attracted much attention in recent years. Its goal is to categorize a collection of unlabeled images into a set of predefined classes for semantic-level image retrieval. Among various image classification methods, many researchers have developed a set of sophisticated models, to represent the spatial context of the local patches in the images, e.g., hidden conditional random fields [2], constellation model [3], etc. Among them, 2-dimensional hidden Markov model (2-D HMM) has attracted much attention as a classic spatial contextual model [4]–[6]. This model can efficiently capture the spatial context among different patches in the images. In more detail, when using 2-D HMM for image categorization, a model is first learned from a training set of images for each image class. Then this learned model can be used to score the probability of an unlabeled image belonging to this class. However, the images in one class usually have large intra-class variance and this variance often leads to the difficulty in constructing a common spatial contextual model for this class. Fig. 1 illustrates an example of this difficulty. The images in the category "car" have many different views in this example, such as top view, side view, front view, and back view. Each view has a different spatial context of their local patches. These differences between the image spatial contexts bring large intra-class variance for this category. As stated above, the traditional 2-D HMM attempts to use a common model to generate all these images with different spatial structures. Therefore, the depictive ability of a single model is too limited to capture large intra-class variance perfectly. Actually, the above problem also exists in many other spatial-contextual models for image categorization, which attempt to use one common generative model to represent one class, such as HCRF [2] and constellation model [3].

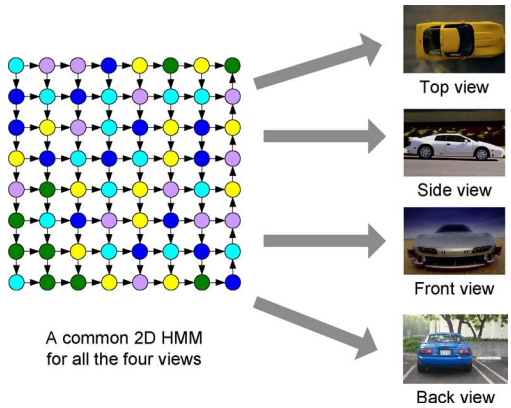


Fig. 1. Using one common 2-D HMM to model the category "car" with large intra-class variance. In this example, four different views in "car" make one model inadequate to capture such an intra-class variance.

To overcome the drawback in the above models, we propose a different kernelized spatial-contextual model for image categorization. This model can better capture the intra-class variance. The underlying motivation is to separate the representation model from the classification model, so that it need not be limited to construct a single model for one class. We will detail this idea in the following section.

As aforementioned, the problem of the traditional 2-D HMM is it attempts to use only one common model to represent an image class with large intra-class variance. Instead of only using

Manuscript received June 14, 2009; revised October 01, 2009 and January 13, 2010; accepted January 21, 2010. First published March 22, 2010; current version published May 14, 2010. A previous short conference version of this paper has been published in [1]. In this paper, we give a more comprehensive description and analysis of the proposed joint appearance and spatial feature representation both theoretically and empirically. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Francesco G. B. De Natale.

G.-J. Qi is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801-2300 USA(e-mail: qi4@illinois.edu).

X.-S. Hua is with the Internet Media Group, Microsoft Research Asia, Beijing 100190, China (e-mail: xshua@microsoft.com).

Y. Rui and H.-J. Zhang are with the Microsoft Advanced Technology Center, Beijing 100190, China (e-mail: yongrui@microsoft.com; hjzhang@microsoft.com).

J. Tang is the School of Computing, National University of Singapore, Singapore 119076 (e-mail: tangjh@comp.nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2046270

one common model, our idea is to construct a set of "prototype" models, each of which captures one "prototype" in this class. For example, for the concept "car" illustrated in Fig. 1, we can construct at least four different "prototype" models to represent these views. By combining these four prototypes, we can obtain a better spatial contextual model for "car". Moreover, for the image classification task, these prototypes can not only capture the intra-class variance but also help to discriminate this concept from the others.

With the above motivation, we are inspired to use the support vector machine (SVM) [7] together with 2-D HMM to model spatial context and discriminate an image class with large variance from other classes in an integrated model. As is well known, SVM has a powerful discrimination ability to find a set of prototype samples which can be used to distinguish different image classes. These samples are called by the support vectors (SVs) in SVM. For the classification task, these SVs contain the complete information for an image class to discriminate itself from other different classes. However, in the traditional manner, these samples in SVM are some feature vectors with fixed length extracted from each image. That is to say, these feature vectors do not preserve any spatial contextual information in them. Thus, we shall develop a model which can not only preserve spatial context of images but also be well embedded into SVM formulation.

Fortunately, when training an SVM model, we do not need to use the original feature model directly. Instead, in its dual formulation, we only need compute the kernel functions between the image representations. In essence, this kernel functions reveal the similarity measure between the images. If we can design such a similarity measure between different spatial contextual models such as 2-D HMMs, we can find a prototype set of spatial-contexts for an image class which can be used to "optimally" distinguish this class from others in terms of maximum marginal principle [7] used in SVM.

Formally, given a set of training images $\{\boldsymbol{x}_i\}_{i=1}^n$ and their associated labels $\{y_i\}_{i=1}^n$, we first learn an individual 2-D HMM $\{\Theta_i\}_{i=1}^n$ from each image. By computing the similarity measure between these models, we can obtain a kernel function $k(\Theta, \Theta_i)$ between two models Θ, Θ_i. Then a prediction function can be learned by SVM as

$$f(\Theta) = \operatorname{sgn}\left\{\sum_{i=1}^{n} y_i \alpha_i k(\Theta, \Theta_i) + b\right\} \tag{1}$$

where $\operatorname{sgn}\{\cdot\}$ is the sign function. Θ and Θ_i are the learned 2-D HMM from the images, α_i and b are the coefficients and bias. This function gives the predicted label for the image associated with model Θ. Those Θ_i's with the associated $\alpha_i \neq 0$ act as support vectors in SVM which can be seen as the prototypes for an image class to discriminate itself from the other categories.

Fig. 2 illustrates an example of such an SVM for the category "car". We can find it contains a set of SVs for the concept "car", each of which represents a certain view. By combining them as (1), they can be used to predict the labels of "car" for the images.

Now the crucial problem of this kernelized spatial contextual model becomes to design a proper similarity measure between different images based on their respective 2-D HMM models.

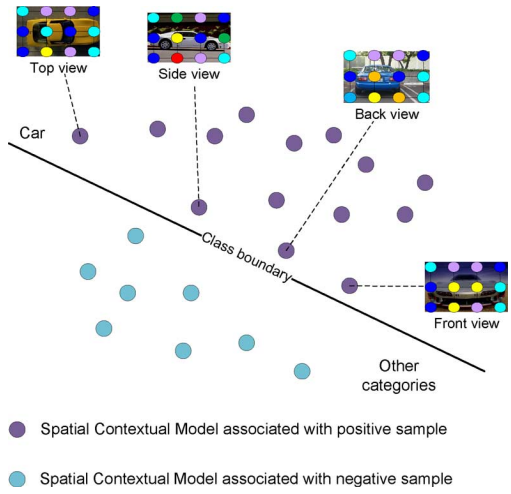


Fig. 2. Example of 2-D HMM.

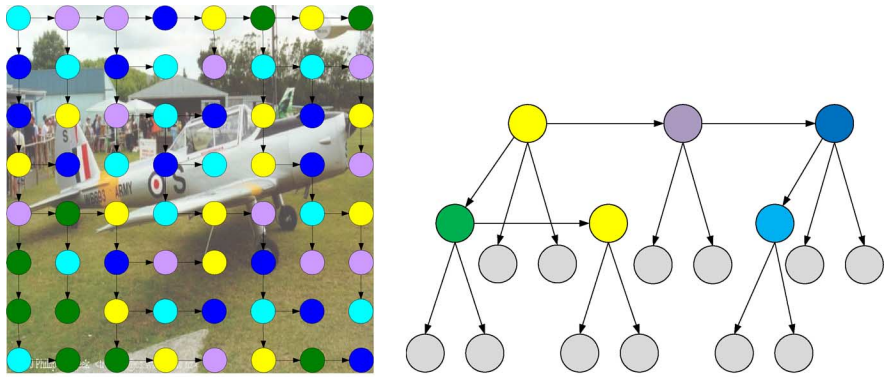
Actually, the image similarities have attracted much attention in recent years. Some researchers attempt to compute the similarities based on image appearances. For example, [8] utilizes the Gaussian mixture models (GMMs) to represent a set of local patches and then computes the similarity between two images via the Fisher kernel between their respective GMMs. Other researchers attempt to compute image similarities according to their semantic distances between their labels [9]. However, all the above algorithms do not consider the spatial context of local patches in the associated images when computing the image similarities, which we believe it is an important factor to discriminate different image categories. Specifically, our goal is to find an image similarity measure between spatial contextual models which jointly considers the appearance-spatial distances between images. As is well known, once such a distance (dissimilarity) is computed, the similarity can be easily obtained. In Section II, we will detail such a distance measure.

II. Two-Dimensional Hidden Markov Model

In Section II-A, we review some related works on two-dimensional hidden Markov model. After that, a statistical model, dependency-tree hidden Markov model (DT-HMM) [10] is introduced to represent spatial structure of an image together with its appearance. To capture the multi-modal features, we propose to extend this model by combining a variety of cues from different feature sources.

A. Survey for Two-Dimensional Hidden Markov Model

Two-dimensional hidden Markov model has been intensively studied, and many researchers have proposed their own models. Li *et al.* [4] extend the traditional one-dimensional HMM into a two-dimensional model by incorporating state dependencies between the neighboring image blocks along both directions of images. However, the computational cost will become impractical with increment of image sizes so that an approximation algorithm is needed to give tractable inferences. After that, they also propose another 2-D HMM to classify the images into different categories and propagate annotations from keywords assigned to those categories [6]. Similarly, Yu *et al.* [5] also pro-



(a) (b)

Fig. 3. (a) Dependency-tree hidden Markov model and (b) a multi-modal variant.

pose to build a spatial HMM for each concept class, and propagate annotations for keywords associated with some specific classes. Their model considers both the vertical and horizontal transitions between hidden states. Recently, Merialdo *et al.* [10] propose an alternative 2-D HMM with a more tractable structure and a significant efficiency can be obtained with an inference on its structure. This model also retains the rich depictive ability to represent the two dimensional dependencies between hidden states on the images. Therefore, we use it into our model to represent an image. But the other 2-D HMMs can be applied to our framework as well.

B. Brief Introduction to Dependency-Tree Hidden Markov Model

DT-HMM is a new 2-D probabilistic modeling approach proposed in [10]. It addresses the complexity of the other modeling approaches such as 2-D HMM [4] while preserving the richness of 2-D representation abilities and having a tractable exact inference procedure.

Similar to that in 2-D HMM, we denote a 2-D observation by $O = \{o_{i,j}, i = 1, \ldots, R, j = 1, \ldots, C\}$, where each $o_{i,j}$ is the feature vector of a block (i, j) in the image. Let there be Q states $\{1, \ldots, Q\}$ and the state of block (i, j) is denoted by $s_{i,j}$. Under the typical dependency assumption in 2-D-HMM, each state $s_{i,j}$ depends on its two neighbors $s_{i-1,j}, s_{i,j-1}$, which usually makes the computation complexity of the learning and inference procedure exponentially grow with the image size in practice [4]. In contrast, the idea of DT-HMM is to assume $s_{i,j}$ only depends on one neighbor at a time. This neighbor may be the horizontal or the vertical one, depending on a random variable $t(i, j)$ with the following distribution:

$$P(t(i,j) = (i-1,j)) = P(t(i,j) = (i,j-1)) = \frac{1}{2}. \quad (2)$$

It is worth noting that for the first row or the first column, $t(i, j)$ has only one valid horizontal or vertical value. $t(1, 1)$ is not defined. So the transition probability distribution can be simplified as

$$P(s_{i,j} \mid s_{i-1,j}, s_{i,j-1}) = \begin{cases} P_V(s_{i,j} \mid s_{i-1,j}), & t(i,j) = (i-1,j) \\ P_H(s_{i,j} \mid s_{i,j-1}), & t(i,j) = (i,j-1) \end{cases} \quad (3)$$

where P_V and P_H are the vertical and horizontal transition probability distributions, respectively. The random variables t for all (i, j) defines a tree-structured dependency over all positions with $(1, 1)$ as the root. Fig. 3(a) illustrates such a dependency tree structure. In terms of computation cost, this structure is highly efficient in inference and learning.

C. Multi-Modal DT-HMM With Multiple Feature Cues

Based on the above DT-HMM, we present how to combine the multiple feature cues into this model. The underlying motivation to combine multiple feature cues is one single feature often cannot capture the complete discriminative differences between the images. For example, as for the "white sand" on the beach and the "snow" in the skiing image, it is not enough to distinguish them merely by the color feature. If the texture features are also incorporated, they can be discriminated into correct classes while the "sand" has the coarser texture and the "snow" has the finer one.

In DT-HMM, given a state $s_{i,j}$, the observation $o_{i,j}$ is generated according to a certain distribution $P(o_{i,j} \mid s_{i,j})$. In this paper, we use GMM as this observation distribution. In the multi-modal setting, the observation $o_{i,j}$ has M feature cues $\{o_{i,j}^k\}_{k=1}^M$ from different sources. We assume these M types of features can be generated independently once the corresponding state $s_{i,j}$ is given, that is

$$\begin{aligned} P\left(\{o_{i,j}^k\}_{k=1}^M \middle| s_{i,j} = q\right) &= \prod_{k=1}^M P\left(o_{i,j}^k \middle| s_{i,j} = q\right) \\ &= \prod_{k=1}^M \sum_{l=1}^N \lambda_{k,l}^q \mathrm{N}\left(o_{i,j}^k \middle| \mu_{k,l}^q, \Sigma_{k,l}^q\right) \end{aligned} \quad (4)$$

where $\lambda_{k,l}^q, \mu_{k,l}^q, \Sigma_{k,l}^q$ is the mixing coefficient, the mean vector, and covariance matrix of lth Gaussian component for the kth modality, respectively, given the current state is q. For simplicity, the covariance matrix is assumed to be diagonal. Fig. 3(b) illustrates such a multi-modal DT-HMM structure. It is worth noting that the independence assumption only holds given hidden states are fixed, and for the whole 2-D observation, such independence assumption does not hold across different modalities, i.e., $P(O^1, \ldots O^M) \neq P(O^1) \cdots P(O^M)$. This

means one feature modality has some statistical dependency on others, so these multiple types of the features can have an effect on each other.

III. Joint Appearance-Spatial Distance Between DT-HMMS

In this section, we will propose how to measure a joint appearance-spatial distance between two images represented by DT-HMM models.

A. Distance Between Models

DT-HMM can be used to jointly encode the appearance and spatial structure. If a proper distance is computed between DT-HMMs, the appearance-spatial discrimination can be measured across the images. From information theory, Kullback-Leibler divergence (KLD) [11] is a natural distance measure between the statistical models.

Specifically, the DT-HMM can be specified completely by the parameter set $\Theta = \{\pi, a^H, a^V, \lambda, \mu, \Sigma\}$, where π is the initial state distribution; a^H, a^V is the horizontal and vertical transition matrix with $a^H_{m,n} = P_H(s_{i,j} = n \,|\, s_{i,j-1} = m), a^V_{m,n} = P_V(s_{i,j} = n \,|\, s_{i-1,j} = m)$; λ, μ, Σ are the parameters for the observation distribution specified in (4). Then the joint distribution of the 2-D observation $O = \{o^k_{i,j}, i = 1, \ldots, R, j = 1, \ldots, C, k = 1, \cdots, M\}$ and state $S = \{s_{i,j}, i = 1, \ldots, R, j = 1, \ldots, C\}$ is

$$\begin{aligned} P(O, S \,|\, \Theta) &= P(O \,|\, S, \Theta) P(S \,|\, \Theta) \\ &= \prod_{i,j} P(o_{i,j} \,|\, s_{i,j}, \Theta) \\ &\quad \times P(s_{i,j} \,|\, s_{i-1,j}, s_{i,j-1}) \end{aligned} \tag{5}$$

and the 2-D observation distribution can be obtained by summarizing S as

$$P(O \,|\, \Theta) = \sum_S P(O, S \,|\, \Theta). \tag{6}$$

Now the KLD between two DT-HMMs $\Theta, \tilde{\Theta}$ is

$$KL(\Theta \,\|\, \tilde{\Theta}) = \int P(O \,|\, \Theta) \log \frac{P(O \,|\, \Theta)}{P(O \,|\, \tilde{\Theta})}. \tag{7}$$

However, there exists no closed-form expression for the KLD between these two DT-HMMs. The most straightforward approach to computing this KLD is to use the Monte Carlo simulation. But that will result in a significant computational cost. In this section, we will present an alternative approximation approach that can be computationally more efficiently than the Monte Carlo approach. This approach is inspired to compute a KLD upper bound between the models [12], [13] by utilizing the following log-sum inequality that has been widely used in information theory [11].

Lemma 1: Given two mixture distributions $f = \sum_{i=1}^L w_i f_i$ and $g = \sum_{i=1}^L v_i g_i$, the KLD between them is upper bounded by

$$KL(f \,\|\, g) \le KL(w \,\|\, v) + \sum_{i=1}^L w_i KL(f_i \,\|\, g_i) \tag{8}$$

where $KL(w \,\|\, v) = \sum_{i=1}^L w_i \log(w_i)/(v_i)$. This inequality directly follows the log-sum inequality (see [11, p. 31]).

This inequality was first used in [14] and in this paper, we will extend it to compute the KLDs between the above DT-HMM models. Let $T(i,j)$ be the sub-tree rooted at position (i,j), and $\beta_{i,j}(q)$ be the probability that the portion of the image is covered by $T(i,j)$ with the state q in position (i,j). Then the whole 2-D observation distribution is

$$P(O \,|\, \Theta) = \sum_{q=1}^Q \pi_q \beta_{1,1}(q). \tag{9}$$

Accordingly, the KLD between two DT-HMMs is then

$$\begin{aligned} KL(\Theta \,\|\, \tilde{\Theta}) &= KL\left(\sum_{q=1}^Q \pi_q \beta_{1,1}(q) \,\|\, \sum_{q=1}^Q \tilde{\pi}_q \tilde{\beta}_{1,1}(q)\right) \\ &\le KL(\pi \,\|\, \tilde{\pi}) \\ &\quad + \sum_{q=1}^Q \pi_q KL(\beta_{1,1}(q) \,\|\, \tilde{\beta}_{1,1}(q)). \end{aligned} \tag{10}$$

The inequality comes from the Lemma 1. The term $KL(\beta_{1,1}(q) \,\|\, \tilde{\beta}_{1,1}(q))$ in the right-hand side can be computed recursively based on an extension of Baum-Welch algorithm by considering the following cases [see Fig. 3(a)].

Case 1: If (i,j) is a leaf in $T(i,j)$ that has no child node [see Fig. 4(a)]:

$$\beta_{i,j}(q) = P(o_{i,j} \,|\, s_{i,j} = q, \Theta). \tag{11}$$

For simplicity of the notation, we denote $\mathrm{N}(o^k_{i,j} \,|\, \mu^q_{k,l}, \Sigma^q_{k,l})$ and $\mathrm{N}(o^k_{i,j} \,|\, \tilde{\mu}^q_{k,l}, \tilde{\Sigma}^q_{k,l})$ by $N^k_{i,j}$ and $\tilde{N}^k_{i,j}$, respectively. Substituting (4) into the above equation, the KLD can be computed as

$$\begin{aligned} &KL(\beta_{i,j}(q) \,\|\, \tilde{\beta}_{i,j}(q)) \\ &\quad = KL\left(\prod_{k=1}^M \sum_{l=1}^N \lambda^q_{k,l} \mathrm{N}^q_{k,l} \,\|\, \prod_{k=1}^M \sum_{l=1}^N \tilde{\lambda}^q_{k,l} \tilde{\mathrm{N}}^q_{k,l}\right) \\ &\quad = \sum_{k=1}^M KL\left(\sum_{l=1}^N \lambda^q_{k,l} \mathrm{N}^q_{k,l} \,\|\, \sum_{l=1}^N \tilde{\lambda}^q_{k,l} \tilde{\mathrm{N}}^q_{k,l}\right) \\ &\quad \le \sum_{k=1}^M \left\{ KL\left(\lambda^q_{k,\cdot} \,\|\, \tilde{\lambda}^q_{k,\cdot}\right) \right. \\ &\qquad \left. + \sum_{l=1}^N \lambda^q_{k,l} KL\left(\mathrm{N}^q_{k,l} \,\|\, \tilde{\mathrm{N}}^q_{k,l}\right) \right\} \end{aligned} \tag{12}$$

where $KL(\lambda^q_{k,\cdot} \,\|\, \tilde{\lambda}^q_{k,\cdot}) = \sum_{l=1}^N \lambda^q_{k,l} \log(\lambda^q_{k,l})/(\tilde{\lambda}^q_{k,l})$. Here, the second equality follows the chain rule for KLD [11] and the inequality comes from the lemma.

Case 2: If (i,j) has only an horizontal successor [see Fig. 4(b)], we have the following recursive equation:

$$\beta_{i,j}(q) = P(o_{i,j} \,|\, s_{i,j} = q, \Theta) \sum_{q'=1}^Q a^H_{q,q'} \beta_{i,j+1}(q') \tag{13}$$

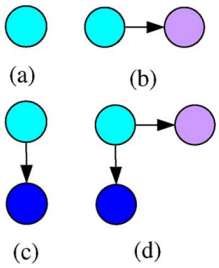


Fig. 4. Four different cases in the DT-HMM structure. (a) Node with no successor. (b) Node with horizontal successor. (c) Node with vertical successor. (d) Node with both horizontal and vertical successor.

thus, we have

$$\begin{aligned} &KL(\beta_{i,j}(q)\,\|\,\tilde{\beta}_{i,j}(q)) \\ &\quad = KL(P(o_{i,j}\,|\,s_{i,j}=q,\Theta)\,\|\,P(o_{i,j}\,|\,s_{i,j}=q,\tilde{\Theta})) \\ &\qquad + KL\left(\sum_{q'=1}^{Q} a^H_{q,q'}\beta_{i,j+1}(q') \,\middle\|\, \sum_{q'=1}^{Q} \tilde{a}^H_{q,q'}\tilde{\beta}_{i,j+1}(q')\right) \\ &\quad \le \sum_{k=1}^{M}\left\{KL\left(\lambda^q_{k,\cdot}\,\middle\|\,\tilde{\lambda}^q_{k,\cdot}\right)\right. \\ &\qquad \left. + \sum_{l=1}^{N}\lambda^q_{k,l}KL\left(\mathrm{N}^q_{k,l}\,\middle\|\,\tilde{\mathrm{N}}^q_{k,l}\right)\right\} + KL\left(a^H_{q,\cdot}\,\|\,\tilde{a}^H_{q,\cdot}\right) \\ &\qquad + \sum_{q'=1}^{Q} a^H_{q,q'}KL(\beta_{i,j+1}(q')\,\|\,\tilde{\beta}_{i,j+1}(q')) \end{aligned} \tag{14}$$

where $KL(a^H_{q,\cdot}\,\|\,\tilde{a}^H_{q,\cdot}) = \sum_{l=1}^{Q} a^H_{q,l}\log(a^H_{q,l})/(\tilde{a}^H_{q,l})$ accounts for the discrimination information of the horizontal spatial structure between the two images. The first equality follows the chain rule for KLD and the inequality comes from the lemma.

Similarly, if (i,j) has only a vertical successor [see Fig. 4(c)], we have

$$\begin{aligned} &KL(\beta_{i,j}(q)\,\|\,\tilde{\beta}_{i,j}(q)) \\ &\quad \le \sum_{k=1}^{M}\left\{KL\left(\lambda^q_{k,\cdot}\,\middle\|\,\tilde{\lambda}^q_{k,\cdot}\right)\right. \\ &\qquad \left. + \sum_{l=1}^{N}\lambda^q_{k,l}KL\left(\mathrm{N}^q_{k,l}\,\middle\|\,\tilde{\mathrm{N}}^q_{k,l}\right)\right\} + KL\left(a^V_{q,\cdot}\,\|\,\tilde{a}^V_{q,\cdot}\right) \\ &\qquad + \sum_{q'=1}^{Q} a^V_{q,q'}KL(\beta_{i+1,j}(q')\,\|\,\tilde{\beta}_{i+1,j}(q')). \end{aligned} \tag{15}$$

Similarly, $KL(a^V_{q,\cdot}\,\|\,\tilde{a}^V_{q,\cdot}) = \sum_{l=1}^{Q} a^V_{q,l}\log(a^V_{q,l})/(\tilde{a}^V_{q,l})$ accounts for the discrimination of the vertical spatial structure between the two images.

Case 3: The last case is that (i,j) has both a horizontal and a vertical successors [see Fig. 4(d)], so we have

$$\begin{aligned} \beta_{i,j}(q) = P(o_{i,j}\,|\,s_{i,j}=q,\Theta)\cdot &\left(\sum_{q'=1}^{Q} a^H_{q,q'}\beta_{i,j+1}(q')\right) \\ \cdot &\left(\sum_{q'=1}^{Q} a^V_{q,q'}\beta_{i+1,j}(q')\right). \end{aligned} \tag{16}$$

Then

$$\begin{aligned} &KL(\beta_{i,j}(q)\,\|\,\tilde{\beta}_{i,j}(q)) \\ &\quad = KL(P(o_{i,j}\,|\,s_{i,j}=q,\Theta)\,\|\,P(o_{i,j}\,|\,s_{i,j}=q,\tilde{\Theta})) \\ &\qquad + KL\left(\sum_{q'=1}^{Q} a^H_{q,q'}\beta_{i,j+1}(q') \,\middle\|\, \sum_{q'=1}^{Q} \tilde{a}^H_{q,q'}\tilde{\beta}_{i,j+1}(q')\right) \\ &\qquad + KL\left(\sum_{q'=1}^{Q} a^V_{q,q'}\beta_{i+1,j}(q') \,\middle\|\, \sum_{q'=1}^{Q} \tilde{a}^V_{q,q'}\tilde{\beta}_{i+1,j}(q')\right) \\ &\quad \le \sum_{k=1}^{M}\left\{KL\left(\lambda^q_{k,\cdot}\,\middle\|\,\tilde{\lambda}^q_{k,\cdot}\right)\right. \\ &\qquad \left. + \sum_{l=1}^{N}\lambda^q_{k,l}KL\left(\mathrm{N}^q_{k,l}\,\middle\|\,\tilde{\mathrm{N}}^q_{k,l}\right)\right\} + KL\left(a^H_{q,\cdot}\,\|\,\tilde{a}^H_{q,\cdot}\right) \\ &\qquad + \sum_{q'=1}^{Q} a^H_{q,q'}KL\left(\beta_{i,j+1}(q')\,\|\,\tilde{\beta}_{i,j+1}(q')\right) \\ &\qquad + KL\left(a^V_{q,\cdot}\,\|\,\tilde{a}^V_{q,\cdot}\right) \\ &\qquad + \sum_{q'=1}^{Q} a^V_{q,q'}KL\left(\beta_{i+1,j}(q')\,\|\,\tilde{\beta}_{i+1,j}(q')\right). \end{aligned} \tag{17}$$

Note that, since DT-HMM has a tree structure, the two subtrees $T(i+1,j)$ and $T(i,j+1)$ have no common nodes. Therefore, the two distributions $(\sum_{q'=1}^{Q} a^H_{q,q'}\beta_{i,j+1}(q'))$ and $(\sum_{q'=1}^{Q} a^V_{q,q'}\beta_{i+1,j}(q'))$ are independent. Thus, in the first equality, we can apply the chain rule for KLD. The inequality still follows the lemma.

Finally, the KLD between the two d-dimensional normal distributions $\mathrm{N}^q_{k,l}, \tilde{\mathrm{N}}^q_{k,l}$ in the above equations has a closed-form expression:

$$\begin{aligned} &KL\left(\mathrm{N}^q_{k,l}\,\middle\|\,\tilde{\mathrm{N}}^q_{k,l}\right) \\ &= \frac{1}{2}\left(\log\frac{|\tilde{\Sigma}^q_{k,l}|}{|\Sigma^q_{k,l}|} + \mathrm{Tr}\left(\left(\tilde{\Sigma}^q_{k,l}\right)^{-1}\Sigma^q_{k,l}\right) + \right. \\ &\qquad \left. \left(\mu^q_{k,l}-\tilde{\mu}^q_{k,l}\right)^T\left(\tilde{\Sigma}^q_{k,l}\right)^{-1}\left(\mu^q_{k,l}-\tilde{\mu}^q_{k,l}\right) - d\right). \end{aligned} \tag{18}$$

B. Further Discussions

There are still two issues that need to be considered when computing the joint distance between DT-HMMs:

1) Once the structure variable t in (2) for DT-HMMs is given, the above KLD is computed with this fixed structure. However, the complete likelihood of DT-HMM given an image is

$$P(O\,|\,\Theta) = \sum_{t} P(O\,|\,t,\Theta)P(t) \tag{19}$$

where the summation is taken over all possible tree structures. Here, all dependency trees are supposed to be equally likely so that $P(t)$ is uniformly distributed. The summation on the right-most term cannot be exhaustively computed by enumerating all possible trees. However, as proven in [10], it can be estimated efficiently by generating only a

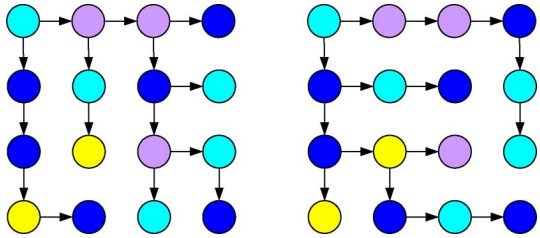


Fig. 5. Two dual DT-HMM structures.

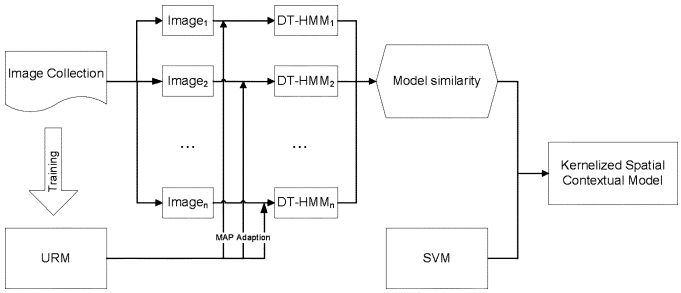


Fig. 6. Adapting each individual DT-HMM from a universal reference model. Such an adaption can give a reasonable correspondence of hidden states between two different DT-HMM.

few trees and averaging over their likelihood. More specifically, the complete likelihood can be effectively computed over two dual trees t and t^τ [10], i.e.,

$$P(O\,|\,\Theta) = \frac{1}{2}\{P(O\,|\,t,\Theta) + P(O\,|\,t^\tau,\Theta^\tau)\} \tag{20}$$

where t^τ is the dual tree of t, defined by replacing horizontal by vertical dependencies and vice versa, except for the boundary constraints. This formulation introduces both horizontal and vertical dependencies for all neighbor pairs in the 2-D observation. Fig. 5 illustrates two DT-HMMs with dual structures. It has been proven in [10] that such a dual approximation has a satisfactory performance compared to the approach by averaging over a large number of trees. Accordingly, the KLD between $\Theta, \tilde{\Theta}$ is

$$\begin{aligned} &KL(\Theta\,||\,\tilde{\Theta}) \\ &\quad = KL\left(\begin{array}{l}\frac{1}{2}\{P(O\,|\,t,\Theta) + P(O\,|\,t^\tau,\Theta^\tau)\}\,||\\ \frac{1}{2}\{P(O\,|\,t,\tilde{\Theta}) + P(O\,|\,t^\tau,\tilde{\Theta}^\tau)\}\end{array}\right) \\ &\quad \le \frac{1}{2}\left\{\begin{array}{l} KL(P(O\,|\,t,\Theta)\,||\,P(O\,|\,t,\tilde{\Theta})) + \\ KL(P(O\,|\,t^\tau,\Theta)\,||\,P(O\,|\,t^\tau,\tilde{\Theta}^\tau))\end{array}\right\} \\ &\quad = \frac{1}{2}\{KL^t(\Theta\,||\,\tilde{\Theta}) + KL^{t^\tau}(\Theta\,||\,\tilde{\Theta})\} \end{aligned} \tag{21}$$

where $KL^t(\Theta\,||\,\tilde{\Theta})$ and $KL^{t^\tau}(\Theta\,||\,\tilde{\Theta})$ are the KLD between Θ and $\tilde{\Theta}$ given the structure t and its dual t^τ, respectively. Here, the above inequality still follows the lemma. From Fig. 5, we can find these two dual structures covers all possible horizontal and vertical spatial structures and thus can give a complete spatial discriminative information between $\Theta, \tilde{\Theta}$.

2) The KLD is not a symmetric measure. We use the following standard symmetric version as the distance measure when implementing the algorithm:

$$D(\Theta\,||\,\tilde{\Theta}) = \frac{1}{2}\{KL(\Theta\,||\,\tilde{\Theta}) + KL(\tilde{\Theta}\,||\,\Theta)\}. \tag{22}$$

Once the symmetric KLD is computed, a kernel can be obtained accordingly. Here, we simply exponentiate the symmetric KLD, i.e.,

$$K(\Theta, \tilde{\Theta}) = \exp\left\{-\frac{D(\Theta\,||\,\tilde{\Theta})}{2\sigma^2}\right\} \tag{23}$$

where σ is the kernel radius. Note that we use an upper bound to approximate the true KL distance between two DT-HMMs; thus, their corresponding kernel matrix from (23) may not be positive-definite. However, there are many solutions to address this problem. For example, in [15], they suggest to compute the smallest eigenvalue of the kernel matrix, and if it is negative, its absolute value can be added to the diagonal of the kernel matrix. This method can be justified as follows. The kernel matrix can be explained intuitively as similarities between images. Adding a positive value to the diagonal only enhances "self-similarities" and it does not affect the similarities among images. Moreover in practice, we have found the obtained kernel matrix often satisfies the positive-definite condition because the computed upper bound of KLD between images is tight enough to the true KLD.

Such a kernel can be applied into a kernel-based classifier. In this paper, we use multi-class SVM [16] for image categorization under the one-versus-the-rest rule: a classifier is learned to separate each class from the rest and the test image is assigned the label of the classifier with one highest score.

3) Here, we analyze the complexity to compute the above distance measure between images. According to the above recursive rules in (10), (12), (14), (15), (17), and (18), the KLD between two DT-HMMs can then be recursively computed in the reverse order, starting from the leaf node until (1, 1). It is not difficult to verify that the computational cost for this upper bound is mainly from computing all the $\beta_{i,j}(q)$, and the computation complexity is $\mathcal{O}(R \cdot C \cdot Q)$ which scales well to 2-D observation size $R \cdot C$. Thus, this distance can be tractably computed.

IV. Adapting DT-HMM From A Universal Reference Model

As stated in Section III, we use an upper bound to approximate the intractable exact KLD between two DT-HMMs. These two models have the same state number Q. However, since they are trained independently on their own images, the correspondence between their respective states may not be in the same order from 1 to Q. Such a disaccord between the states in the two models can lead to an upper bound that is not tight enough. To obtain a tighter bound, as illustrated in Fig. 6, we can first train a *universal reference model* (URM) from referential images, e.g., background images or images from a training set. Then given an image, its DT-HMM can be adapted from this URM. Since the models are all adapted from this URM, the states will have a

reasonable correspondence between two models. Thus, the obtained upper bound will be much tighter than that computed from the independently-trained models.

In this paper, the standard maximum *a posteriori* (MAP) technique [17] is use to adapt the DT-HMM. Formally, given the parameters of the URM Θ^{URM} and 2-D observation O of the new image, we estimate the new DT-HMM Θ. We use Θ^{URM} as the initial parameter. As suggested in [17], the standard expectation-maximization (EM) algorithm is then applied to update Θ repeatedly until convergence except for the mean vector of GMMs, i.e.,

$$\mu_{k,l}^{q} \leftarrow \alpha\mu_{k,l}^{q} + (1-\alpha) \cdot \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} o_{i,j} P\left(s_{i,j}=q, m_{i,j}^{q,k}=l \,|\, O,\Theta\right)}{\sum_{i=1}^{R}\sum_{j=1}^{C} P\left(s_{i,j}=q, m_{i,j}^{q,k}=l \,|\, O,\Theta\right)} \quad (24)$$

where $m_{i,j}^{q,k}$ indicates the mixture component for the kth modality given the state is q at position (i,j), and α is the weighting factor giving the bias between the previous estimate and the current one. We will set α to be 0.7 in the experiment. The update rules for all the other parameters follow the EM algorithm.

V. Experiments

In this section, we will conduct extensive experiments to compare the proposed approach against the original DT-HMM proposed in [10] and the other two best representative kernel methods: multiple-instance learning via embedded instance selection (MILES) [18] and spatial pyramid matching (SPM) [19]. MILES is a bag-of-word algorithm which does not model the spatial context of local patches. Its source code is publicly available at http://john.cs.olemiss.edu/ychen/data/MILES.zip. On the other hand, similarly, SPM represents the state-of-the-art kernel-based image categorization algorithm, where it models the spatial context by using the geometric correspondence to match the spatial layout of the local features.

For all the three approaches, there are algorithmic parameters that need to be determined. To ensure a fair comparison, all the parameters in all three approaches are determined by a twofold cross-validation process on training set. The reported results are from the best set of parameters in the three approaches. The comparison is conducted on two widely used data sets, one grayscale (the scene data set) and one color (the Corel data set).

A. Natural Scene Image Classification

The first data set is one of the most complete scene category dataset in the literature [19], [20]. It is composed of 15 grayscale scene categories: 13 were provided by Li *et al.* in [20], and the other two were collected by Lazebnik *et al.* in [19].

For the experiment, we follow the same setup in SPM [19]. Namely, we randomly select 100 images per class for training and the rest for testing. All the images from the training set are used to train a URM model. All experiments are repeated ten times with different training and testing images, and the average of per-class classification accuracy is reported. The experiments reported in SPM [19] are conducted with the SIFT descriptor. For the sake of fair comparison, comparison between MILSE, SPM and the proposed approach is using SIFT, too. Specifically, the 128-dimemsional SIFT descriptor is processed by principal component analysis (PCA) to reduce its dimensionality to 50. Regarding training the DT-HMM model, all the images in each category are used to train an individual model for this category base on maximum likelihood criterion. In testing phase, prediction is made by assigning to each testing sample the category with maximum probability given by the associated model.

TABLE I
Average Classification Accuracies (%) for the Three Algorithms on 15 Scene Dataset

Algorithm	Average accuracy
DT-HMM [10]	67.1
Fei-Fei et al. [20]	65.2
MILES [18]	75.4
SPM [19]	81.4
The proposed approach	**87.0**

To ensure meaningful comparison, we use extra care when extracting features, trying to maximize the strength for each approach. For SPM, DT-HMM, and the proposed algorithm, the SIFT are computed in 16-by-16 pixel patches over a grid with spacing of 8 pixels. As for MILES, we follow its original way of extracting features [18] to ensure its best performance. First, salience regions are identified using the approach introduced in [21], which detects regions that are salient over both location and scale. Each salient region is cropped from image and also scaled to an image patch with a size of 16-by-16 pixel, from which the features (SIFT and CM) are extracted. We also report and compare the results of the algorithm proposed by Li *et al.* [20].

The results are shown in Table I and are consistent with our analysis in the paper: the proposed approach has obtained the best performance of all the four algorithms. Compared to the original DT-HMM algorithm, it improves its accuracy by 29.7% from 67.1% to 87.0%. In most natural scene categories, it is usually difficult for the traditional DT-HMM to find a common spatial structure to describe and represent all the images of this category. In contrast, the superior performance of the proposed approach probably comes from its ability to capture the large intra-class variance in the natural scene image dataset. On the other hand, the proposed approach also outperforms MILES because it takes spatial structure into account, as well as SPM because it follows the *least commitment principle* and the distance measure is based on an integrated joint appearance-spatial feature. Therefore, from this experiment, the proposed approach has been proven to outperform the traditional spatial contextual model like DT-HMM but also outperform the other competitive kernel-based image categorization method such as SPM and MILES. In the next section, we will conduct experiments to compare these algorithms on a complex hybrid scene/object image dataset.

Finally, we compare the times spent on computing KL distances between DT-HMM by the typical Monte Carlo method and the proposed approach under the above experimental setting. We conduct the experiments on a computer with 3 GHz CUP and 1 GB memory. On average, Monte Carlo method uses 150 ms for computing KL distance between a pair of DT-HMM while the proposed approach uses 20 ms. It nearly accelerates

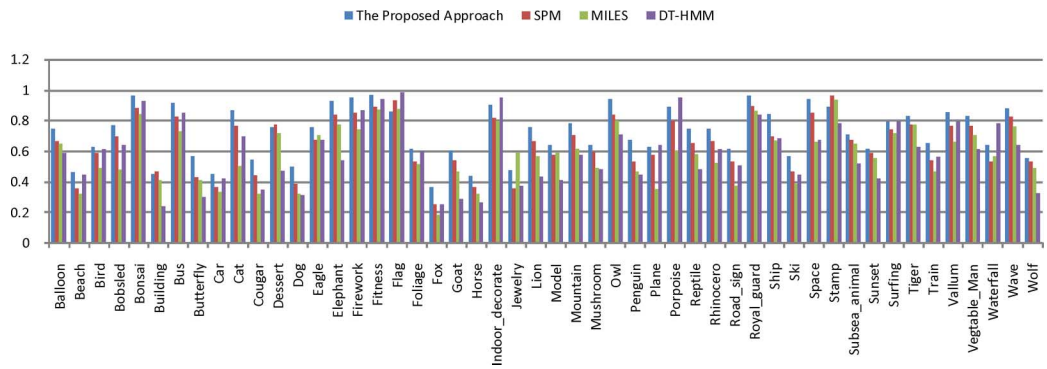


Fig. 7. Classification accuracy over all the 50 categories on the modality of color moment.

the computation by one order of magnitude. We also compare the prediction times made by MILES, SPM, and the proposed approach which are all kernel-based algorithms. They spend 10.7 ms, 2.1 ms, and 6.2 ms on average to predict a sample. We can find SPM uses the least time to predict since it uses a code-book-based method, while MILES and the proposed algorithm spend a little more time to compute the kernel function between the testing sample and the support vectors. But as for the proposed method, the compensation is a better classification accuracy as shown in Table I.

B. Hybrid Scene/Object Image Classification

1) Dataset and Experiment Setup: The second data set is a hybrid object/scene data set from the Corel image collection. Different from many other widely-used Corel datasets which are probably the most widely used [22], this Corel dataset consists of 50 semantically diverse categories with 100 images per category. In these 50 categories, 37 of them contain a certain target object for recognition; the other 13 categories have images for natural scenery. It is a challenging data set because: 1) it has many variations in illumination, occlusion, viewpoint change, cluttered backgrounds, etc.; 2) for the object categories, an image often contains more than one targeted objects and the objects usually do not locate at the center of the image; 3) for the natural scene categories, the images in the same categories often vary significantly in appearance, spatial layout, and lighting conditions.

During the experiment, the images in each category are randomly split into five parts of equal size. We successively use each of the five parts as testing set, and the others are used for training. The URM model is trained from all the images in the training set. The average classification accuracies over these five different testing set is then reported for evaluation.

Because the Corel data set is a color image set, we extract the color moments (CM) features in addition to the SIFT features. Before extracting CM, it is advantageous to convert the images into a perceptual-sensible color space, such as CIE Luv space. The first to third moments of each band are computed, respectively, on the local patches of the image. We therefore have 9-D CM features.

2) Performance Comparison With Previous Methods: Table II shows the average classification accuracies for the three algorithms over all the 50 image categories. Similar

TABLE II
AVERAGE CLASSIFICATION ACCURACIES (%) FOR MILES, SPM, AND THE PROPOSED ALGORITHM ON TWO MODAL FEATURES CM AND SIFT

Algorithm	CM	SIFT
DT-HMM [10]	60.7	41.8
MILES [18]	58.6	43.3
SPM [19]	65.1	49.4
The proposed approach	**72.4**	**56.0**

observations can be obtained as in Section V-A. The proposed approach outperforms the traditional DT-HMM algorithm. It improves its accuracy by 19.3% and 34.0% on color moment and SIFT modalities, respectively. This result is consistent with our expectation that the proposed approach can capture large intra-class variances in the image categories due to different object views or various spatial layout in different natural scenes. To show it, we illustrate the top 10 "support vectors" with the largest coefficients for "car" in Fig. 9. We can find that many different views of this object have been included into this prototype set of spatial contextual models. It reveals the underlying reasons that lead to the better performance of the proposed approach.

On the other hand, among all the kernel-based algorithms, SPM outperforms MILES because it takes spatial structure into account. The proposed approach outperforms SPM because it follows the *least commitment principle* and the distance measure is based on an integrated joint appearance-spatial feature. Furthermore, CM outperforms SIFT, as compared to the other researchers' results [23]. However, it is true that the results usually depend on the specific dataset used for testing the algorithm, where different features have different performance on different datasets [24], [25].

We also illustrate the classification accuracy of all the 50 categories on the color moment and SIFT modalities in Figs. 7 and 8, respectively. The proposed approach obtains the best performance over 41 and 33 categories on color moment and SIFT modality, respectively.

3) Classification Results on Multi-Modal Features: Finally we also do experiment by combining these SIFT and CM feature cues by using the multi-modal DT-HMM proposed in Section II-C. In Fig. 10, we illustrate the confusion matrix on this Corel collection with combined features. As we can see, the classification accuracy is further improved to be 77.3%. This result justifies such a multi-modal strategy can improve

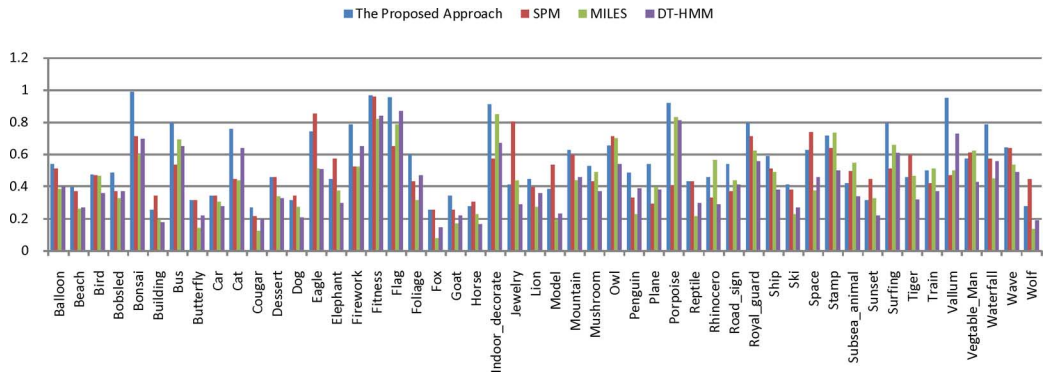


Fig. 8. Classification accuracy over all the 50 categories on the modality of the SIFT.

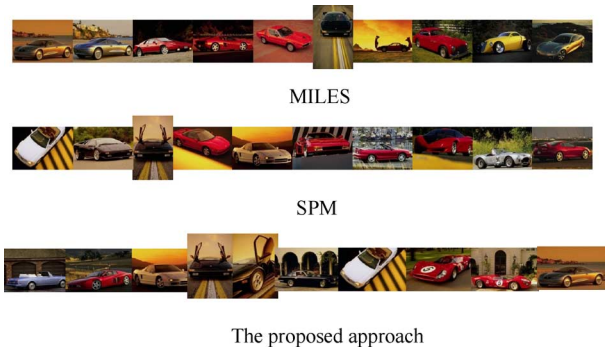


Fig. 9. Top ten images acting as support vectors with the largest coefficients in MILES, SPM, and the proposed approach, which are shown under each image. From this figure, we can find in the proposed approach, among these support vectors, different views of "car" with large intra-class variance have been included to form a prototype set of the spatial context for the category "car".

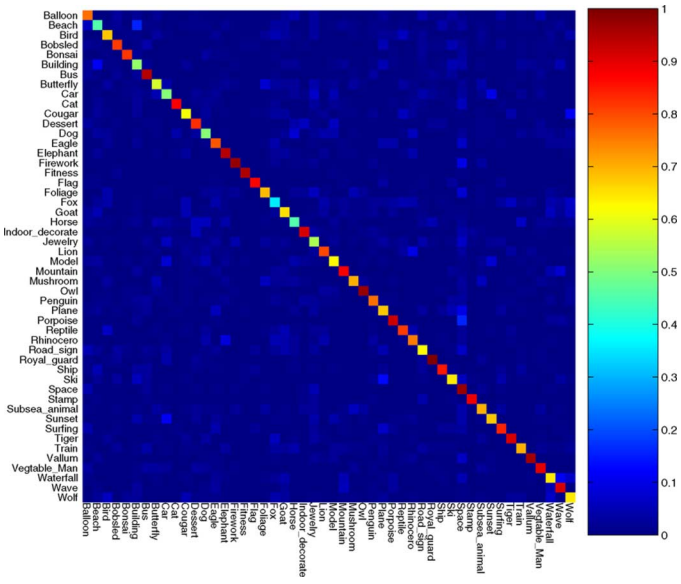


Fig. 10. Confusion matrix on hybrid scene/object data set from Corel collection with the multiple feature cues. The average classification accuracy on these 50 concepts is 77.3%.

the discrimination ability compared to the single-modal one (72.4% on CM modality and 56.0% on SIFT modality).

VI. Conclusion

In this paper, we propose a kernelized spatial contextual model to jointly representing and classifying images in an integrated framework. In contrast to the traditional 2-D HMM which attempts to use one single model to represent the images in one class with large intra-class variance, we construct a prototype set of images by embedding the spatial contexts into the kernelization algorithm. Moreover, these prototypes contains the complete information to distinguish one class of images from the others. Therefore, such an algorithm combines the advantages of rich representation ability of spatial contextual models as well as the powerful classification ability of kernel method.

To embed the spatial model into kernel method, we propose to design a similarity measure between them. This similarity measure is computed by computing the model distances between different images. The distance measures used in most existing approaches either ignored the spatial structures or used them in a separate step. To address these difficulties, in this paper, we proposed a new distance measure that integrates joint appearance-spatial image features. In addition, multiple modal features can be incorporated into this distance measure to help improve the discrimination ability across multiple features. We further proposed an efficient algorithm to compute this distance. Its upper bound can be further tightened by adapting a universal reference model into individual probabilistic models. Extensive experiments on two image data sets demonstrate that the proposed approach outperforms the state-of-the-art approaches in both scene and object images.

References

[1] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, Z.-J. Zha, and H.-J. Zhang, "A joint appearance-spatial distance for kernel-based image categorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[2] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Proc. Advances in Neural Information Processing System*, 2004.

[3] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2003, pp. 264–271.

[4] J. Li, A. Najmi, and R. M. Gray, "Image classification by a two dimensional hidden Markov model," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 517–533, Feb. 2000.

[5] F. Yu and H. lp, "Automatic semantic annotation of images using spatial hidden Markov model," in *Proc. IEEE ICME*, 2006.

[6] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, Oct. 2003.
[7] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[8] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE CVPR*, 2007.
[9] J. Yu, J. Amores, N. Sebe, P. Radeva, and Q. Tian, "Distance learning for similarity estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 451–462, Mar. 2008.
[10] B. Merialdo, J. Jiten, E. Galmar, and B. Huet, "A new approach to probabilistic image modeling with multimensional hidden Markov models," in *Proc. 4th Int. Workshop Adaptive Multimedia Retrieval*, 2006.
[11] T. Cover and J. Thomas, *Elements of Information Theory*, ser. Wiley Series in Telecommunications, 2nd ed. New York: Wiley, 2006.
[12] M. Do and M. Vetterli, "Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden Markov models," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 517–527, Dec. 2002.
[13] P. Liu, F. K. Soong, and J.-L. Zhou, "Divergence-based similarity measure for spoken document retrieval," in *Proc. IEEE ICASSP*, 2007.
[14] Y. Singer and M. K. Warmuth, "Batch and on-line parameter estimation of Gaussian mixtures based on the joint entropy," in *Proc. NIPS*, 1998.
[15] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE CVPR*, 2006.
[16] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm.
[17] J.-L. Gauvain and C.-H. Lee, "Maximum a posteri estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
[18] Y. Chen, J. Bi, and Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.
[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE CVPR*, 2006.
[20] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE CVPR*, 2005.
[21] T. Kadir and M. Brady, "Scale, saliency and image description," *Int. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, 2001.
[22] J. Tang and P. Lewis, "A study of quality issues for image auto-annotation with the corel data-set," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 384–389, Mar. 2007.
[23] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Local ensemble kernel learning for object category recognition," in *Proc. IEEE CVPR*, 2007.
[24] J. He, S.-F. Chang, and L. Xie, "Fast kernel learning for spatial pyramid matching," in *Proc. IEEE CVPR*, 2008.
[25] K. V. de Sande, T. Gevers, and C. Snoek, "Evaluation of color descriptors for object and scene recognition," in *Proc. IEEE CVPR*, 2008.

Guo-Jun Qi received the B.E. degree in automation from the University of Science and Technology of China, Hefei, Anhui, China, in 2005.

He is currently with the Image Formation and Processing group at the University of Illinois at Urbana-Champaign. From November 2008 to January 2009, he worked as a research assistant in the School of Computing at the National University of Singapore. From July 2006 to November 2008, he was a research intern in the Internet Media Group at Microsoft Research Asia. His research interests include machine learning and data mining, especially with their applications into multimedia computing and analysis.

Mr. Qi was the winner of the Best Paper Award at the 15th ACM International Conference on Multimedia, Augsburg, Germany, in 2007.

Xian-Sheng Hua (M'05) received the B.S. and Ph.D. degrees from Peking University, Beijing, China, in 1996 and 2001, respectively, both in applied mathematics.

When he was in Peking University, his major research interests were in the areas of image processing and multimedia watermarking. Since 2001, he has been with Microsoft Research Asia, Beijing, where he is currently a Lead Researcher with the internet media group. His current interests are in the areas of video content analysis, multimedia search, management, authoring, sharing, and advertising.

Yong Rui (F'10) received the B.S. degree from Southeast University, Nanjing, China, the M.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree from the University of Illinois at Urbana-Champaign (UIUC).

He serves as Director of Strategy of Microsoft China R&D (CRD) Group, Beijing.

He is a Senior Member of ACM. He is an Associate Editor of the *ACM Transactions on Multimedia Computing, Communication and Applications* (TOMCCAP), the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He was an Editor of the *ACM/Springer Multimedia Systems Journal* and the *International Journal of Multimedia Tools and Applications* from 2005–2007. He also serves on the Advisory Board of the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING.

Jinhui Tang (M'08) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, in July 2003 and July 2008, respectively.

He is currently a postdoctoral research fellow in the School of Computing, National University of Singapore. From June 2006 to February 2007, he worked as a research intern in the Internet Media group at Microsoft Research Asia, Beijing, China. From February 2008 to May 2008, he worked as a research intern in the School of Computing, National University of Singapore. His current research interests include content-based image retrieval, video content analysis, and pattern recognition.

Hong-Jiang Zhang (M'91–SM'97–F'03) received the B.S. degree from Zhengzhou University, Henan, China, in 1982 and the Ph.D. degree from the Technical University of Denmark, Lyngby, in 1991, both in electrical engineering.

From 1992 to 1995, he was with the Institute of Systems Science, National University of Singapore, where he led several projects in video and image content analysis and retrieval and computer vision. From 1995 to 1999, he was a research manager at Hewlett-Packard Labs, Palo Alto, CA, where he was responsible for research and development in the areas of multimedia management and intelligent image processing. In 1999, he joined Microsoft Research, where he is currently the Managing Director of Advanced Technology Center in Beijing, China. He has coauthored/coedited four books, over 350 papers and book chapters, numerous special issues of international journals on image and video processing, content-based media retrieval, and computer vision, as well as over 60 granted patents.

Dr. Zhang is a Fellow of ACM. He currently serves on the editorial board of the PROCEEDINGS OF THE IEEE.