A Joint Appearance-Spatial Distance for Kernel-Based Image Categorization

[†]Guo-Jun Qi, [‡]Xian-Sheng Hua, [‡]Yong Rui, [†]Jinhui Tang, [†] Zheng-Jun Zha, [‡]Hong-Jiang Zhang
[†]MOE-Microsoft Key Laboratory of Multimedia Computing and Communication
& Department of Automation, University of Science and Technology of China
Huang Shan Road, No.4, Hefei, Anhui, 230027, China
{qgj, jhtang, zzjun}@mail.ustc.edu.cn
[‡]Microsoft Research Asia
Microsoft Research Asia, 49 Zhichun Road, Beijing, 100080, China
{xshua, yongrui, hjzhang}@microsoft.com

Abstract

The goal of image categorization is to classify a collection of unlabeled images into a set of predefined classes to support semantic-level image retrieval. The distance measures used in most existing approaches either ignored the spatial structures or used them in a separate step. As a result, these distance measures achieved only limited success. To address these difficulties, in this paper, we propose a new distance measure that integrates joint appearance-spatial image features. Such a distance measure is computed as an upper bound of an information-theoretic discrimination, and can be computed efficiently in a recursive formulation that scales well to image size. In addition, the upper bound approximation can be further tightened via adaption learning from a universal reference model. Extensive experiments on two widely-used data sets show that the proposed approach significantly outperforms the state-of-the-art approaches.

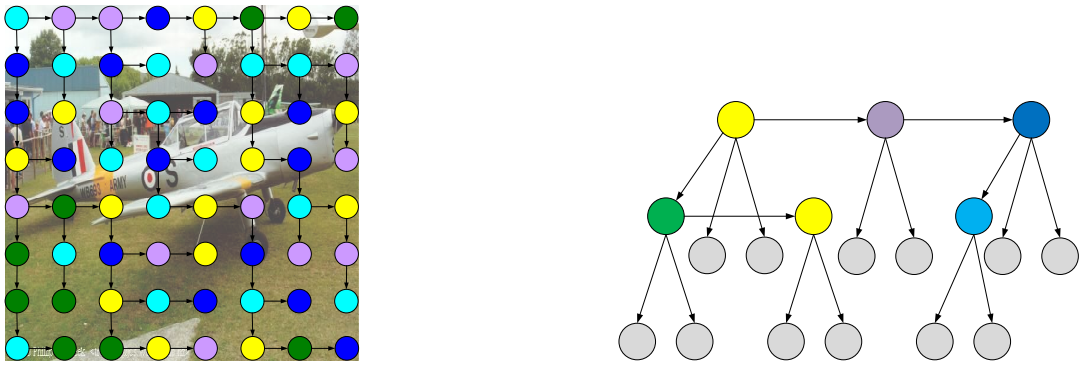
1. Introduction

The goal of image categorization is to classify a collection of unlabeled images into a set of predefined classes for semantic-level image retrieval. Although much effort has been made to improve the categorization accuracy, it is still an open research problem in the computer vision community. The major difficulties come from both image *appearance variations*, e.g., material differences, background clutters and lighting changes, and complex *spatial variations*, e.g., different structures of object parts, occlusions, and changes in viewpoints. Here, the appearance and spatial structure can be represented by using local patches [6] and their spatial layout in the images [5][10].

In the context of using local features to do image categorization, there exist many different learning/classification methods. Among them, the kernel-based methods attracted most attention and represent some of the best performance in the field [2][8][16][18][12]. The core of these kernel based methods is to design an image distance measure to compute the (dis)similarity between two images. Ideally, the distance measure should capture both the appearance and spatial structure of the underlying images.

Over the past decade, the kernel-based methods for image categorization evolved through two major paradigms, characterized by their different distance measures. The *first paradigm* is the bag-of-feature approaches that represent an image as an orderless collection of local features [2][8][16][18]. The distance (dissimilarity) between two images is measured by the two bag-of-feature sets. For example, in [2] the images are first embedded into a bag-level feature space and the image distance is computed based on the bag-of-feature representation of image. These algorithms ignore the spatial structure of the local features, which limits their description capability, especially in capturing the shape of objects and the spatial coherence between local features.

To overcome the difficulties in the first paradigm, the *second paradigm* approaches take advantage of the spatial coherence between local features to compute the distance. It attempts to directly "match" the geometric correspondence between image patches on a global scale by aggregating statistics of local patches over some fixed sub-regions [10]. Following such a research direction, a pyramid matching algorithm was further developed and achieved better results than the first-paradigm approaches [10]. While this approach takes into account of the spatial structures, they have the following limitations. (1) The distance measure is based on comparing images' absolute spatial layouts of the predefined visual words; and more importantly, (2) These



(a) Dependency-Tree Hidden Markov Model (b) A multi-modal variant

Figure 1. Dependency-Tree Hidden Markov Model (a) and a multi-modal variant (b)

approaches handle the appearance and spatial features in two separate steps. The second-step's spatial matching between the visual words is highly influenced by the accuracy of the first-step appearance matching through visual vocabulary. It is difficult for this two-step approach to achieve global optimum. The appearance matching errors incurred in the first step will propagate to the second step.

To overcome the difficulties in the second paradigm, in this paper, we propose the *third paradigm*, where the spatial structural are not only taken into account, but jointly with the appearance features in an integrated framework. This paradigm follows the *Least Commitment Principle*, advocated by David Marr [13]. Instead of matching appearance features and spatial structures in two separate steps, the proposed approach addresses the image matching problem via a joint distance measure without a preprocessed visual vocabulary. Specifically, the proposed approach computes the distance between two images by computing an information-theoretic distance between two statistical models that encode the appearance and spatial distribution of the images. The upper bound for the distance can be efficiently computed by a recursive procedure which scales well to the size of the images. An even tighter bound of this distance can be obtained by a Maximum A Posteriori (MAP) adaption scheme, i.e., each statistical model for the image is adapted from a Universal Reference Model (URM), which is constructed from a collection of referential images. Once such an upper bound is obtained, a kernel function can be obtained in a kernel-based classifier, such as Support Vector Machine (SVM), for image categorization.

The rest of this paper is organized as follows. In Section 2, we introduce the probabilistic models for images. In Section 3, an effective information theoretic distance is proposed to measure the discriminative differences between the image probabilistic models based on their joint appearance-spatial information. To obtain a tighter bound for this distance, an adaption scheme is developed in section 4 so that all the underlying probabilistic models are adapted from a URM. In section 5, extensive experiments on two widely-used data sets show that the proposed approach achieves significantly better performance than the existing state-of-the-art approaches. We give concluding remarks in Section 6.

2. Dependency-Tree Hidden Markov Model

In section 2.1, a statistical model, dependency-tree hidden Markov model (DT-HMM) [14] is introduced to represent the appearance and spatial structure of an image. After that, we propose to extend this model to capture the multi-modal features by combining a variety of cues from different feature sources.

2.1. A Brief Introduction to Dependency-Tree Hidden Markov Model

DT-HMM is a new 2D probabilistic modeling approach proposed in [14]. It addresses the complexity of the other modeling approaches such as 2D HMM [11][17] while preserving the richness of 2D representation abilities and having a tractable exact inference procedure.

Similar to that in 2D HMM, we denote a 2D observation by $O = \{o_{i,j}, i = 1, \cdots, R, j = 1, \cdots, C\}$, where each $o_{i,j}$ is the feature vector of a block (i, j) in the image. Let there be Q states $\{1, \cdots, Q\}$ and the state of block (i, j) is denoted by $s_{i,j}$. Under the typical dependency assumption in 2D-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 [11]. 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:

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

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

$$P(s_{i,j}|s_{i-1,j}, s_{i,j-1}) = \begin{cases} P_V(s_{i,j}|s_{i-1,j}), & t(i,j) = (i-1,j) \\ P_H(s_{i,j}|s_{i,j-1}), & t(i,j) = (i,j-1) \end{cases} \tag{2}$$

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. Figure 1(a) illustrates such a dependency tree structure. In terms of computation cost, this structure is highly efficient in inference and learning.

2.2. A 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 $P(o_{i,j}|s_{i,j})$. In this paper, we use Gaussian Mixture Model (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

$$\begin{aligned} &P(\{o_{i,j}^k\}_{k=1}^M|s_{i,j} = q) = \textstyle\prod_{k=1}^M P(o_{i,j}^k|s_{i,j} = q) \\ &= \textstyle\prod_{k=1}^M \sum_{l=1}^N \lambda_{k,l}^q \mathrm{N}(o_{i,j}^k|\mu_{k,l}^q, \Sigma_{k,l}^q) \end{aligned} \tag{3}$$

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. Figure 1(b) illustrates such a multi-modal DT-HMM structure. It is worth of noting that the independence assumption only holds given hidden states are fixed and for the whole 2D observation such independence assumption does not hold across different modalities, i.e. $P(O^1, \cdots O^M) \neq P(O^1) \cdots P(O^M)$. This means one feature modality has some statistical dependency on others, so these multiple types of the features can make effect on each other.

3. A 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.

3.1. 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) [3] 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 $a_{m,n}^H = P_H(s_{i,j} = n|s_{i,j-1} = m), a_{m,n}^V = P_H(s_{i,j} = n|s_{i-1,j} = m)$; λ, μ, Σ are the parameters for the observation distribution specified in Eqn. 3. Then the joint distribution of the 2D observation $O = \{o_{i,j}^k, i = 1, \cdots, R, j = 1, \cdots, C, k = 1, \cdots, M\}$ and state $S = \{s_{i,j}, i = 1, \cdots, R, j = 1, \cdots, C\}$ is

$$\begin{aligned} &P(O, S|\Theta) = P(O|S, \Theta)P(S|\Theta) \\ &= \textstyle\prod_{i,j} P(o_{i,j}|s_{i,j}, \Theta)P(s_{i,j}|s_{i-1,j}, s_{i,j-1}) \end{aligned} \tag{4}$$

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

$$P(O|\Theta) = \sum\nolimits_S P(O, S|\Theta) \tag{5}$$

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

$$D_{KL}\left(\Theta||\tilde{\Theta}\right) = \int P(O|\Theta) \log \frac{P(O|\Theta)}{P(O|\tilde{\Theta})} \tag{6}$$

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 by computing an upper bound of KLD between the models[4].

The approximation is motivated from the following lemma [15] that is based on the log-sum inequality [3]:

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*

$$D_{KL}(f||g) \leq D_{KL}(w||v) + \sum\nolimits_{i=1}^L w_i D_{KL}(f_i||g_i) \tag{7}$$

where $D_{KL}(w||v) = \sum_{i=1}^L w_i \log \frac{w_i}{v_i}$. This inequality directly follows the log-sum inequality (see pp. 31 of [3]).

Given this lemma, the KLD between DT-HMMs can be computed. 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 2D observation distribution is

$$P(O|\Theta) = \sum_{q=1}^{Q} \pi_q \beta_{1,1}(q) \tag{8}$$

Accordingly, the KLD between two DT-HMMs is then

$$\begin{aligned} &D_{KL}\left(\Theta||\tilde{\Theta}\right) = D_{KL}\Big(\textstyle\sum_{q=1}^{Q} \pi_q \beta_{1,1}(q) || \sum_{q=1}^{Q} \tilde{\pi}_q \tilde{\beta}_{1,1}(q)\Big) \\ &\leq D_{KL}\left(\pi||\tilde{\pi}\right) + \textstyle\sum_{q=1}^{Q} \pi_q D_{KL}\left(\beta_{1,1}(q)||\tilde{\beta}_{1,1}(q)\right) \end{aligned} \tag{9}$$

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

Case 1 If (i,j) is a leaf in $T(i,j)$ that has no child node:

$$\beta_{i,j}(q) = P(o_{i,j}|s_{i,j} = q) \tag{10}$$

For simplicity of the notation, we denote $\mathrm{N}(o_{i,j}^k|\mu_{k,l}^q, \Sigma_{k,l}^q)$ and $\mathrm{N}(o_{i,j}^k|\tilde{\mu}_{k,l}^q, \tilde{\Sigma}_{k,l}^q)$ by $N_{i,j}^k$ and $\tilde{N}_{i,j}^k$, respectively. Substituting Eqn. 3 into the above equation, the KLD can be computed as

$$\begin{aligned} &D_{KL}\left(\beta_{i,j}(q)||\tilde{\beta}_{i,j}(q)\right) \\ &= D_{KL}\left(\textstyle\prod_{k=1}^{M}\sum_{l=1}^{N} \lambda_{k,l}^q \mathrm{N}_{k,l}^q || \prod_{k=1}^{M}\sum_{l=1}^{N} \tilde{\lambda}_{k,l}^q \tilde{\mathrm{N}}_{k,l}^q\right) \\ &= \textstyle\sum_{k=1}^{M} D_{KL}\left(\sum_{l=1}^{N} \lambda_{k,l}^q \mathrm{N}_{k,l}^q || \sum_{l=1}^{N} \tilde{\lambda}_{k,l}^q \tilde{\mathrm{N}}_{k,l}^q\right) \\ &\leq \sum_{k=1}^{M}\left\{D_{KL}\left(\lambda_{k,\cdot}^q||\tilde{\lambda}_{k,\cdot}^q\right) + \sum_{l=1}^{N} \lambda_{k,l}^q D_{KL}\left(\mathrm{N}_{k,l}^q||\tilde{\mathrm{N}}_{k,l}^q\right)\right\} \end{aligned} \tag{11}$$

where $D_{KL}\left(\lambda_{k,\cdot}^q||\tilde{\lambda}_{k,\cdot}^q\right) = \sum_{l=1}^{N} \lambda_{k,l}^q \log \frac{\lambda_{k,l}^q}{\tilde{\lambda}_{k,l}^q}$. Here, the second equality follows the chain rule for KLD [3] and the inequality comes from the lemma.

Case 2 If (i,j) has only an horizontal successor, we have the following recursive equation:

$$\beta_{i,j}(q) = P(o_{i,j}|s_{i,j} = q) \textstyle\sum_{q'=1}^{Q} a_{q,q'}^H \beta_{i,j+1}(q') \tag{12}$$

thus we have

$$\begin{aligned} &D_{KL}\left(\beta_{i,j}(q)||\tilde{\beta}_{i,j}(q)\right) \\ &= D_{KL}\left(P(o_{i,j}|s_{i,j} = q, \Theta)||P(o_{i,j}|s_{i,j} = q, \tilde{\Theta})\right) \\ &+ D_{KL}\left(\textstyle\sum_{q'=1}^{Q} a_{q,q'}^H \beta_{i,j+1}(q') || \sum_{q'=1}^{Q} \tilde{a}_{q,q'}^H \tilde{\beta}_{i,j+1}(q')\right) \\ &\leq \sum_{k=1}^{M}\left\{D_{KL}\left(\lambda_{k,\cdot}^q||\tilde{\lambda}_{k,\cdot}^q\right) + \sum_{l=1}^{N} \lambda_{k,l}^q D_{KL}\left(\mathrm{N}_{k,l}^q||\tilde{\mathrm{N}}_{k,l}^q\right)\right\} \\ &+ D_{KL}\left(a_{q,\cdot}^H||\tilde{a}_{q,\cdot}^H\right) + \sum_{q'=1}^{Q} a_{q,q'}^H D_{KL}\left(\beta_{i,j+1}(q')||\tilde{\beta}_{i,j+1}(q')\right) \end{aligned} \tag{13}$$

where $D_{KL}\left(a_{q,\cdot}^H||\tilde{a}_{q,\cdot}^H\right) = \sum_{l=1}^{Q} a_{q,l}^H \log \frac{a_{q,l}^H}{\tilde{a}_{q,l}^H}$ 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, we have

$$\begin{aligned} &D_{KL}\left(\beta_{i,j}(q)||\tilde{\beta}_{i,j}(q)\right) \\ &\leq \sum_{k=1}^{M}\left\{D_{KL}\left(\lambda_{k,\cdot}^q||\tilde{\lambda}_{k,\cdot}^q\right) + \sum_{l=1}^{N} \lambda_{k,l}^q D_{KL}\left(\mathrm{N}_{k,l}^q||\tilde{\mathrm{N}}_{k,l}^q\right)\right\} \\ &+ D_{KL}\left(a_{q,\cdot}^V||\tilde{a}_{q,\cdot}^V\right) + \sum_{q'=1}^{Q} a_{q,q'}^V D_{KL}\left(\beta_{i+1,j}(q')||\tilde{\beta}_{i+1,j}(q')\right) \end{aligned} \tag{14}$$

Similarly, $D_{KL}\left(a_{q,\cdot}^V||\tilde{a}_{q,\cdot}^V\right) = \sum_{l=1}^{Q} a_{q,l}^V \log \frac{a_{q,l}^V}{\tilde{a}_{q,l}^V}$ 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, so we have

$$\begin{aligned} &\beta_{i,j}(q) = P(o_{i,j}|s_{i,j} = q, \Theta) \cdot \left(\textstyle\sum_{q'=1}^{Q} a_{q,q'}^H \beta_{i,j+1}(q')\right) \\ &\cdot \left(\textstyle\sum_{q'=1}^{Q} a_{q,q'}^V \beta_{i+1,j}(q')\right) \end{aligned} \tag{15}$$

Then

$$\begin{aligned} &D_{KL}\left(\beta_{i,j}(q)||\tilde{\beta}_{i,j}(q)\right) \\ &= D_{KL}\left(P(o_{i,j}|s_{i,j} = q, \Theta)||P(o_{i,j}|s_{i,j} = q, \tilde{\Theta})\right) \\ &+ D_{KL}\left(\textstyle\sum_{q'=1}^{Q} a_{q,q'}^H \beta_{i,j+1}(q') || \sum_{q'=1}^{Q} \tilde{a}_{q,q'}^H \tilde{\beta}_{i,j+1}(q')\right) \\ &+ D_{KL}\left(\textstyle\sum_{q'=1}^{Q} a_{q,q'}^V \beta_{i+1,j}(q') || \sum_{q'=1}^{Q} \tilde{a}_{q,q'}^V \tilde{\beta}_{i+1,j}(q')\right) \\ &\leq \sum_{k=1}^{M}\left\{D_{KL}\left(\lambda_{k,\cdot}^q||\tilde{\lambda}_{k,\cdot}^q\right) + \sum_{l=1}^{N} \lambda_{k,l}^q D_{KL}\left(\mathrm{N}_{k,l}^q||\tilde{\mathrm{N}}_{k,l}^q\right)\right\} \\ &+ D_{KL}\left(a_{q,\cdot}^H||\tilde{a}_{q,\cdot}^H\right) + \sum_{q'=1}^{Q} a_{q,q'}^H D_{KL}\left(\beta_{i,j+1}(q')||\tilde{\beta}_{i,j+1}(q')\right) \\ &+ D_{KL}\left(a_{q,\cdot}^V||\tilde{a}_{q,\cdot}^V\right) + \sum_{q'=1}^{Q} a_{q,q'}^V D_{KL}\left(\beta_{i+1,j}(q')||\tilde{\beta}_{i+1,j}(q')\right) \end{aligned} \tag{16}$$

Note that, since DT-HMM has a tree structure, the two sub trees $T(i+1,j)$ and $T(i,j+1)$ have no common nodes. Therefore the two distributions $\left(\sum_{q'=1}^{Q} a_{q,q'}^H \beta_{i,j+1}(q')\right)$ and $\left(\sum_{q'=1}^{Q} a_{q,q'}^V \beta_{i+1,j}(q')\right)$ 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 $\mathrm{N}_{k,l}^q, \tilde{\mathrm{N}}_{k,l}^q$ in the above equations has a closed-

form expression:

$$D_{KL}\left(\mathrm{N}_{k,l}^{q}||\tilde{\mathrm{N}}_{k,l}^{q}\right) = \frac{1}{2}\left(\begin{array}{c}\log\frac{|\tilde{\Sigma}_{k,l}^{q}|}{|\Sigma_{k,l}^{q}|} + Tr\left(\left(\tilde{\Sigma}_{k,l}^{q}\right)^{-1}\Sigma_{k,l}^{q}\right) + \\ \left(\mu_{k,l}^{q}-\tilde{\mu}_{k,l}^{q}\right)^{T}\left(\tilde{\Sigma}_{k,l}^{q}\right)^{-1}\left(\mu_{k,l}^{q}-\tilde{\mu}_{k,l}^{q}\right) - d\end{array}\right) \tag{17}$$

Now, according to the above recursive rules in Eqn. 9 11 13 14 16 17, 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 $\mathcal{O}\left(R\cdot C\cdot Q\right)$ which scales well to 2D observation size $R\cdot C$.

3.2. Implementation issues

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 Eqn. 1 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

$$P(O|\Theta) = \sum\nolimits_{t} P(O|t,\Theta)P(t) \tag{18}$$

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 [14], it can be estimated efficiently by generating only a few trees and averaging over their likelihood. More specifically, the complete likelihood can be effectively computed over two dual trees t and t^{τ} [14], i.e.,

$$P(O|\Theta) = \frac{1}{2}\left\{P(O|t,\Theta) + P(O|t^{\tau},\Theta^{\tau})\right\} \tag{19}$$

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 2D observation. [14] discusses this dual structure in detail. It has been proven in [14] 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

$$\begin{aligned} & D_{KL}\left(\Theta||\tilde{\Theta}\right) \\ &= D_{KL}\left(\begin{array}{c}\frac{1}{2}\left\{P(O|t,\Theta)+P(O|t^{\tau},\Theta^{\tau})\right\}|| \\ \frac{1}{2}\left\{P(O|t,\tilde{\Theta})+P(O|t^{\tau},\tilde{\Theta}^{\tau})\right\}\end{array}\right) \\ &\leq \frac{1}{2}\left\{\begin{array}{c}D_{KL}\left(P(O|t,\Theta)||P(O|t,\tilde{\Theta})\right)+ \\ D_{KL}\left(P(O|t^{\tau},\Theta)||P(O|t^{\tau},\tilde{\Theta}^{\tau})\right)\end{array}\right\} \\ &= \frac{1}{2}\left\{D_{KL}^{t}\left(\Theta||\tilde{\Theta}\right)+D_{KL}^{t^{\tau}}\left(\Theta||\tilde{\Theta}\right)\right\} \end{aligned} \tag{20}$$

where $D_{KL}^{t}\left(\Theta||\tilde{\Theta}\right)$ and $D_{KL}^{t^{\tau}}\left(\Theta||\tilde{\Theta}\right)$ are the KLD between given the structure t and its dual t^{τ}, respectively . Here, the above inequality still follows the lemma. From figure 3, 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

$$D\left(\Theta||\tilde{\Theta}\right) = \frac{1}{2}\left\{D_{KL}\left(\Theta||\tilde{\Theta}\right)+D_{KL}\left(\tilde{\Theta}||\Theta\right)\right\} \tag{21}$$

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

$$K(\Theta,\tilde{\Theta}) = \exp\left\{-\frac{D\left(\Theta||\tilde{\Theta}\right)}{2\sigma^{2}}\right\} \tag{22}$$

where σ is the kernel radius. We summarize the algorithm for constructing this kernel in Alg. 1.

Such a kernel can be applied into a kernel-based classifier. In this paper, we use multi-class Support Vector Machine (SVM) [1] 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.

4. Adapting DT-HMM from a Universal Reference Model

As stated in Section 3, 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

Algorithm 1 Compute the kernel $K\left(\Theta||\tilde{\Theta}\right)$ in Eqn. 22

1: Given the structure t, compute the upper bound for the KLD distance $D^t_{KL}\left(\Theta||\tilde{\Theta}\right)$ according to the recursive Eqn. 9 11 13 14 16 17.

2: Given the dual structure t^τ, compute the dual KLD $D^{t^\tau}_{KL}\left(\Theta||\tilde{\Theta}\right)$ as step 1;

3: Compute the KLD by averaging over structure t and t^τ its dual according to Eqn. 20

$$D_{KL}\left(\Theta||\tilde{\Theta}\right) \approx \frac{1}{2}\left\{D^t_{KL}\left(\Theta||\tilde{\Theta}\right) + D^{t^\tau}_{KL}\left(\Theta||\tilde{\Theta}\right)\right\}$$

4: Repeat from step 1 to step 3, compute $D_{KL}\left(\tilde{\Theta}||\Theta\right)$ and then the symmetric KLD as Eqn. 21

$$D\left(\Theta||\tilde{\Theta}\right) = \frac{1}{2}\left\{D_{KL}\left(\Theta||\tilde{\Theta}\right) + D_{KL}\left(\tilde{\Theta}||\Theta\right)\right\}$$

5: Compute the kernel $K\left(\Theta||\tilde{\Theta}\right)$ as Eqn. 22

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

is not tight enough. To obtain a tighter bound, 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 [7] is use to adapt the DT-HMM. Formally, given the parameters of the URM Θ^{URM} and 2D observation O of the new image, we estimate the new DT-HMM Θ. We use Θ^{URM} as the initial parameter. As suggested in [7], the standard expectation-maximization (EM) algorithm is then applied to update Θ repeatedly until convergence except for the mean vector of GMMs, i.e.

$$\mu^q_{k,l} \leftarrow \alpha\mu^q_{k,l} + (1-\alpha)\cdot\frac{\sum_{i=1}^{R}\sum_{j=1}^{C} o_{i,j}P(s_{i,j}=q, m^{q,k}_{i,j}=l|O,\Theta)}{\sum_{i=1}^{R}\sum_{j=1}^{C} P(s_{i,j}=q, m^{q,k}_{i,j}=l|O,\Theta)} \quad (23)$$

where $m^{q,k}_{i,j}$ indicates the mixture component for 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.

5. Experiments

Multiple-Instance Learning via Embedded Instance Selection (MILES) [2] is one of the best approaches in the first paradigm, outperforming many other bag-of-word algorithms for image categorization. Its source code is publicly available at http://john.cs.olemiss.edu/ ychen/data/MILES.zip. Similarly, Spatial Pyramid Matching (SPM) [10] represents the state-of-the-art in the second paradigm, where it uses the geometric correspondence to match the spatial layout of the local features.

In this section, we will conduct extensive experiments to compare the proposed approach against the two best representatives from the first and second paradigms: MILES and SPM. For all the three approaches, there are algorithmic parameters 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 gray-scale (the scene data set) and one color (the Corel data set).

5.1. The scene data set

The first data set is one of the most complete scene category dataset in the literature [5][10]. It is composed of fifteen grayscale scene categories: thirteen were provided by Li et al. in [5], and the other two were collected by Lazebnik et al. in [10].

For the experiment, we follow the same setup in SPM [10]. Namely, we randomly select 100 images per class for training and the rest for testing. 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 [10] 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 principle component analysis (PCA) to reduce its dimensionality to 50.

To ensure meaningful comparison, we use extra care when extracting features, trying to maximize the strength for each approach. For SPM and the proposed algorithm, the SIFT are computed in a 16-by-16 pixel patches over a grid with spacing of 8 pixels. As for MILES, we follow its original way of extracting features [2] to ensure its best performance. First, salience regions are identified using the approach introduced in [9], which detects regions that are salient over both location and scale. Each salient region is

Algorithm	Average accuracy
Fei-Fei et al. [5]	65.2
MILES [2]	75.4
SPM [10]	81.4
The proposed approach	**87.0**

Table 1. The average classification accuracies (%) for the three algorithms on fifteen scene dataset

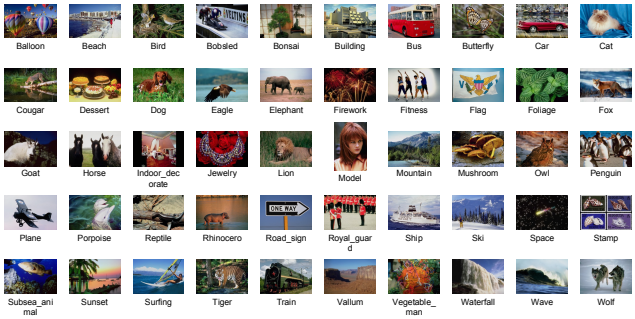


Figure 2. Some example images for 50 category Corel data set

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.

The results are shown in Table 1 and are consistent with our analysis in the paper: 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.

5.2. The Corel data set

The second data set is the Corel data set, which is probably the most widely used in image categorization [2]. It 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. Figure 2 illustrates some example images for this data set. 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 5 parts of equal size. We successively use each of the five parts as testing set, and the others are used for training. 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-dimensional CM features.

Algorithm	CM	SIFT
MILES [2]	58.6	43.3
SPM [10]	65.1	49.4
The proposed approach	**72.4**	**56.0**

Table 2. The average classification accuracies (%) for MILES, SPM and the proposed algorithm on two modal features CM and SIFT.

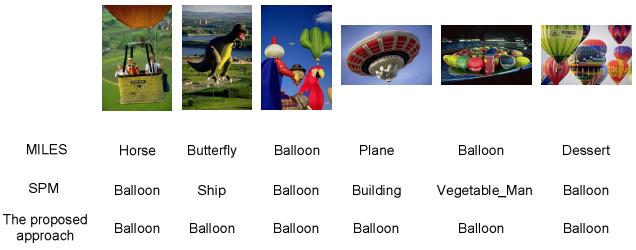


Figure 3. Some classification results on the image category "Balloon" for MILES, SPM and the proposed approach.

Table 2 shows the average classification accuracies for the three algorithms over all the 50 image categories. Similar observations can be obtained as in Section 5.1: 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, which is consistent with other researchers' results that color is an important feature [12].

To illustrate the strength and weakness of the three paradigms, we selected 6 images in the "Balloon" category in Figure 3. It also shows the classification results based on CM feature by using MILES, SPM and the proposed approach. We can see that the bag-of-feature approach MILES misclassifies four "Balloon" images into other categories because it only captures the appearance of the local patches (e.g., the color feature) without considering their spatial configuration. For example, the sixth image has similar color appearance as the category "Dessert", and MILES mistakes it for "Dessert". For the second-paradigm approaches, SPM misclassifies three "Balloon" images. As illustrated from these six images, the "Balloon" category have images with very different appearance. As aforementioned in section 1, since these varying appearance features are scattered in the whole feature space, they can lead to improper visual words that are indistinguishable from other categories. Therefore, the second-paradigm's *spatial*

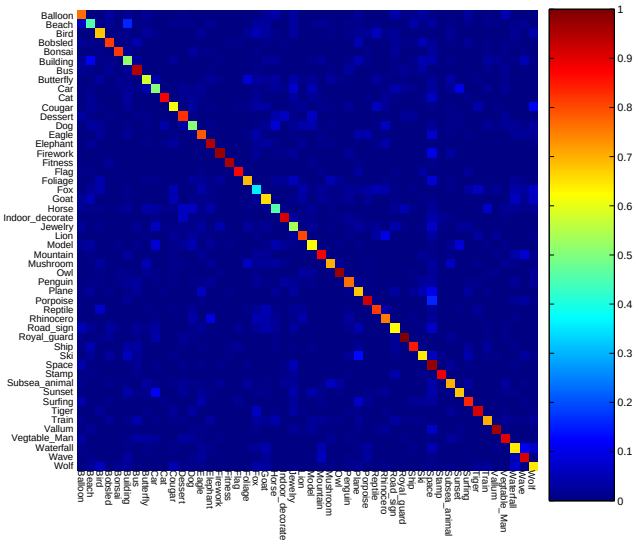


Figure 4. Confusion matrix on hybrid scene/object data set from Corel collection with the multiple feature cues. The average classification accuracy on these 50 concepts is 77.3%.

matching step is frustrated by such a visual word vocabulary, and gives poor results. In contrast, the proposed algorithm overcomes the drawbacks in the first and second paradigms by proposing the joint distances integrating both the appearance and the spatial features. As a result, it classifies the "Balloon" images correctly.

Finally we also do experiment by combining these SIFT and CM feature cues by using the multi-modal DT-HMM proposed in section 2.2. In figure 4, 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 the discrimination ability compared to the single-modal one (72.4% on CM modality and 56.0% on SIFT modality).

6. Conclusion

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. We further proposed an efficient algorithm to compute this distance. Its upper bound can be tightened by adapting a universal reference model into individual probabilistic models. Extensive experiments on two widely-used data sets demonstrate that the proposed approach outperforms the state-of-the-art approaches in both scene and object images.

References

[1] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. 5

[2] Y. Chen, J. Bi, and Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transaction on Pattern Analysis and Machine Learning*, 28(12):1931–1947, 2006. 1, 6, 7

[3] T. Cover and J. Thomas. *Elements of information theory, second edition*. Wiley Series in Telecommunications, John Wiley and Sons, New York, 2006. 3, 4

[4] M. Do and M. Vetterli. Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden markov models. *IEEE Transaction on Multimedia*, (4):517–527, 2002. 3

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of IEEE CVPR*, 2005. 1, 6, 7

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 264–271, 2003. 1

[7] J.-L. Gauvain and C.-H. Lee. Maximum a posteri estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transaction on Speech and Audio Processing*, (2):291–298, 1994. 6

[8] K. Grauman and T. Darrell. Pyramid match kernels: discriminative classification with sets of image features. In *Proc. of IEEE ICCV*, 2005. 1

[9] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001. 6

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE CVPR*, 2006. 1, 6, 7

[11] J. Li, A. Najmi, and R. M. Gray. Image classification by a two dimensional hidden markov model. *IEEE Transaction on Signal Processing*, 48(2):517–533, February 2000. 2

[12] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Local ensemble kernel learning for object category recognition. In *Proc. of IEEE CVPR*, 2007. 1, 7

[13] D. Marr. *Vision*. W. H. Freeman and Company, 1982. 2

[14] B. Merialdo, J. Jiten, E. Galmar, and B. Huet. A new approach to probabilistic image modeling with multimensional hidden markov models. In *Proc. of 4th International Workshop on Adaptive Multimedia Retrieval*, 2006. 2, 5

[15] Y. Singer and M. K. Warmuth. Batch and on-line parameter estimation of gaussian mixtures based on the joint entropy. In *Proc. of NIPS*, 1998. 3

[16] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proc. of IEEE ICCV*, 2003. 1

[17] F. Yu and H. lp. Automatic semantic annotation of images using spatial hidden markov model. In *Proc. of IEEE ICME*, 2006. 2

[18] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: discriminative nearest neighbor classification for visual category recognition. In *Proc. of IEEE CVPR*, 2006. 1