Image Categorization by Learning with Context and Consistency

Zhiwu Lu and Horace H.S. Ip
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
lzhiwu2@student.cityu.edu.hk, cship@cityu.edu.hk

Abstract

This paper presents a novel semi-supervised learning method which can make use of intra-image semantic context and inter-image cluster consistency for image categorization with less labeled data. The image representation is first formed with the visual keywords generated by clustering all the blocks that we divide images into. The 2D spatial Markov chain model is then proposed to capture the semantic context across these keywords within an image. To develop a graph-based semi-supervised learning approach to image categorization, we incorporate the intra-image semantic context into a kind of spatial Markov kernel which can be used as the affinity matrix of a graph. Instead of constructing a complete graph, we resort to a k-nearest neighbor graph for label propagation with cluster consistency. To the best of our knowledge, this is the first application of kernel methods and 2D Markov models simultaneously to image categorization. Experiments on the Corel and histological image databases demonstrate that the proposed method can achieve superior results.

1. Introduction

Image categorization refers to the labeling of images into one of some predefined categories. Though this is usually not a very difficult task for humans, it has been proven to be extremely challenging for machines owing to variable and sometimes uncontrolled imaging conditions as well as complex and hard-to-describe objects in an image. Despite many previous efforts to resolve this problem, there are still two issues in image categorization. One is how to reduce the gap between the low-level visual features extracted from images by machines and the high-level semantics of images. The other is how to utilize the large amount of unlabeled data in applications such as image retrieval since the manual labeling of images is too costly.

To handle the first issue in image categorization, we can consider a semantic intermediate representation for each image, which has been shown effective in [3, 5, 14]. The bag-of-words methods such as probabilistic latent semantic analysis (PLSA) [7] and latent Dirichlet allocation (LDA) [1] are examples of automatically learning image semantics with such an intermediate representation based on visual keywords. However, since all the regions within an image are assumed to be independently drawn from a generation probability distribution, this kind of methods ignore the spatial structure of the image.

Hence, we propose a spatial Markov chain (SMC) model based on the image representation with visual keywords to capture the spatial structure of an image. The formation of visual keywords is achieved through clustering all the blocks that we divide images into. Unlike hidden Markov models [15], the visual feature vectors of blocks are no longer considered by our SMC model, which have been used for generating the visual keywords. Our SMC model is a 2D generalization of the Markov chain by employing a second-order neighborhood system on the regular grid and assuming the conditional independence of vertical and horizontal transitions between states. Although some other 2D extensions of Markov chain have also been proposed such as pseudo 2D Markov model [9] and Markov mesh random field [4], they may not simultaneously consider the vertical and horizontal transitions of states in a tractable solution, or may not capture the local relationship between blocks with rows or columns of states as the calculation elements.

Moreover, to handle the second issue in image categorization, we can resort to semi-supervised learning that is able to make use of both labeled and unlabeled data. In the literature, typical methods include co-training [2] and graph-based semi-supervised learning [20, 22]. In this paper, we will focus on graph-based semi-supervised learning for image categorization, due to the success of applying it to many fields. Although the graph is at the heart of these graph-based methods, its construction has not been studied extensively. More concretely, the Gaussian function is usually adopted to calculate the edge weights of the graph, and the choice of the free parameter (i.e. variance) in this function will affect the results significantly.

For graph construction with the intra-image semantic context, we then propose a spatial Markov kernel (SMK) which can be used as the affinity matrix of a graph. An

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

SMC model is first estimated for each image, and then the kernel is defined to measure how close two SMC models are. Instead of constructing a complete graph, we make use of a k-nearest neighbor (k-NN) graph [6] for label propagation with cluster consistency, i.e., we let each image absorb a fraction of label information from its k-nearest neighbors and retain some label information of its initial state in each propagation step. To the best of our knowledge, this is the first application of kernel methods and 2D Markov models simultaneously to image categorization, although there exist other related works that have combined kernel methods with 1D HMM for protein classification [8] or with Gaussian mixture model for image categorization [13]. Moreover, our method is different from learning with local and global consistency [22] based on a complete graph. Although the idea behind our method is also label propagation in neighborhoods [20], we construct the k-NN graph using our SMK, instead of semi-definite quadratic programming that follows the linear neighborhood assumption.

The remainder of this paper is organized as follows. Section 2 gives a brief review of some previous works that are closely related to this paper. In Section 3, an SMC model is proposed to capture the semantic context across visual keywords within an image. In Section 4, this intra-image semantic context is first incorporated into a kind of SMK and a k-NN graph is then constructed based on this kernel for label propagation with cluster consistency. In Section 5, the proposed method is evaluated on the Corel and histological image databases. Finally, Section 6 then gives the conclusions drawn from our experimental results.

2. Related Work

2.1. Semantic Intermediate Representation

In the context of image categorization, one direct strategy is to classify images using some low-level visual features such as color and texture. This approach considers each category as an individual object [17, 19], which is usually applied to classify only a small number of categories such as indoor versus outdoor or city versus landscape. To reduce the gap between low-level and high-level image processing, we can follow another more effective strategy that adopts a semantic intermediate representation [3, 5, 14] for each image before image categorization and then matches the scene/object model with the perception we humans have. With such a representation, we are then able to classify a larger number of categories.

The bag-of-words methods such as probabilistic latent semantic analysis (PLSA) [7] and latent Dirichlet allocation (LDA) [1] are examples of automatically learning image semantics using the latter strategy. Each image is first represented by the frequency of visual keywords, which are learnt through dividing all the images into regions and then applying a clustering algorithm on the visual feature vectors of all the regions (i.e., each cluster is a keyword). A mixture of latent topics is further used to model each image, and the topics are learnt as multinomial distributions of visual keywords. It should be stressed that these bag-of-words methods ignore the spatial structure of the image, since all the regions of an image are assumed to be independently drawn from the mixture of latent topics.

2.2. Hidden Markov Models

Most natural or histological images have an inherent layered structure. For example, for the beach scene, there are three horizontal regions (layers), starting from the bottom of the image: sand, water, and sky. To learn the spatial structure of the image, a popular type of probabilistic graphical models, i.e. the Markov models such as hidden Markov model (HMM) [15] and Markov random field [16], have been widely applied in image semantic analysis. Most previous works on HMM mainly captured directly the transitions of low-level visual features extracted from different regions (or blocks) across the image or across different resolutions (or scales) of the image [10, 11].

Recently, HMM has been applied to model the transitions of high level semantic labels across an image [21]. In these applications of HMM to capture the spatial dependencies between semantic labels, each image has to be first divided into equivalent blocks on a regular grid so that the spatial position can be characterized for each block. Through defining the notion of "past" as what we have observed in a row-wise raster scan on the regular grid, a novel 2D spatial hidden Markov model (SHMM) has been proposed to capture the spatial dependencies between semantic labels within an image. Though this SHMM can be used to make automatic semantic annotation for each block in a test image, the semantic labels of all the blocks from the training images have to been provided to train the model, which is a challenging task for a large database.

2.3. Graph-Based Semi-Supervised Learning

In recent years, a prominent achievement in the semi-supervised learning area is the development of a graph-based semi-supervised learning strategy, which models the whole data set as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here, $\mathcal{V}$ is the vertex set, which is equivalent to the data set, and $\mathcal{E}$ is the edge set. Associated with each edge $e_{ij} \in \mathcal{E}$ is a nonnegative weight $w_{ij} \geq 0$ reflecting how similar two data points i and j are. The basic idea behind the graph-based semi-supervised learning methods is label propagation on graphs with the cluster consistency [22], which can be stated that two data points on the same structure (cluster or manifold) are likely to have the same class label.

Though the graph is at the heart of these graph-based methods, its construction has not been studied extensively

[20]. More concretely, most of them adopt a Gaussian function to calculate the edge weights of the graph, and the choice of the free parameter (i.e. variance) in this function will affect the classification results significantly. Here, it should be noted that the selection of this parameter with less labeled data is a challenging task.

3. Semantic Context Analysis

In order to apply Markov models to image categorization, we have to first generate the image representation based on visual keywords. That is, all the images are first divided into equivalent blocks on a regular grid, and then some representative properties are extracted for each block by incorporating the color and texture features. Based on the extracted blocks, a vocabulary of visual keywords $\mathcal{S} = \{s_i\}_{i=1}^{M}$ is then generated to exploit the content similarities of blocks. With this universal vocabulary $\mathcal{S}$, we can then represent each image as a 2D sequence of visual keywords by a row-wise raster scan on the regular grid. In this representation, a visual keyword is automatically attached to each block in the image. More formally, an image with $X \times Y$ blocks can be denoted as $Q = q_{11}q_{12}...q_{1Y}q_{21}q_{22}...q_{2Y}...q_{XY}$, where $q_{xy} \in \mathcal{S}$ $(1 \le x \le X,\ 1 \le y \le Y)$ is the visual keyword of block (x, y) in the image. This 2D sequence representation is particularly suitable for the Markov models.

We now present our spatial Markov chain (SMC) model as follows. Let $Q_{x,y}$ denote the sequence of states (i.e. keywords) from block $(1, 1)$ to block (x, y) in a row-wise raster scan on the regular grid. The defining property for an SMC model can then be formulated as

$$P(q_{x,y}|Q_{x,y-1}) = P(q_{x,y}|q_{x,y-1}, q_{x-1,y}). \tag{1}$$

That is, the probability of a state $q_{x,y}$ given its previous state sequence $Q_{x,y-1}$ is equal to the probability of $q_{x,y}$ given its contextual states–$q_{x,y-1}$ for its preceding block and $q_{x-1,y}$ for its upper block. We further assume that the vertical and horizontal state transitions are conditional independent, i.e., the right-hand part of the above formula can be calculated as the product of these two parts. More formally, it can be given in detail as follows

$$P(q_{x,y}|q_{x,y-1}, q_{x-1,y}) = \begin{cases} P(q_{1,1}), & x = y = 1; \\ P(q_{1,y}|q_{1,y-1}), & y > x = 1; \\ P(q_{x,1}|q_{x-1,1}), & x > y = 1; \\ P(q_{x,y}|q_{x,y-1})P(q_{x,y}|q_{x-1,y}), & x, y > 1. \end{cases} \tag{2}$$

The underlying idea is that the state of a block only depends on the states of two previously neighbor blocks in a row-wise raster scan on a non-symmetric half plane. This assumption of our SMC model differs dramatically from that of Markov mesh random field models [4]. The second-order neighborhood system used in our Markov model is further graphically illustrated in Figure 1.

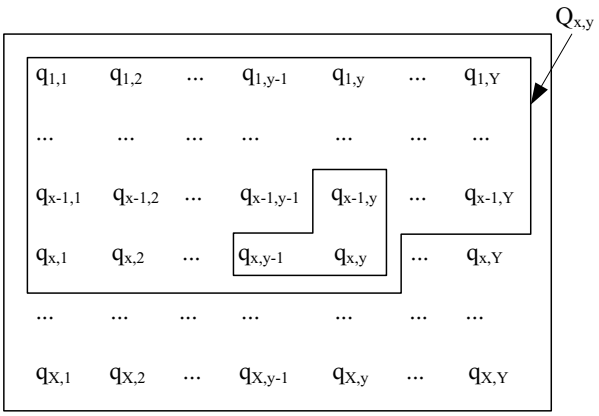


Figure 1. The second-order neighborhood system used in the proposed SMC model.

Elements of the proposed 2D Markov model can now be formally defined as follows:

(1) $\pi = \{\pi_i\}$ is the collection of initial state distribution, where $\pi_i = P(q_{1,1} = s_i)$.

(2) The horizontal state transition matrix is $H = \{h_{i,j} : s_i, s_j \in \mathcal{S},\ 1 \le i, j \le M\}$, where $h_{i,j}$ is defined as $P(q_{x,y} = s_j|q_{x,y-1} = s_i)$, i.e., the transition from state s_i to state s_j along the horizontal direction.

(3) The vertical state transition matrix is $V = \{v_{i,j} : s_i, s_j \in \mathcal{S},\ 1 \le i, j \le M\}$, where $v_{i,j}$ is defined as $P(q_{x,y} = s_j|q_{x-1,y} = s_i)$, i.e., the transition from state s_i to state s_j along the vertical direction.

For convenience, the compact notation $\lambda = \{\pi, H, V\}$ is used to indicate the complete parameter set of an SMC model. In the following, we will discuss two basic problems associated with the SMC model and also present our solutions in detail.

The first problem is to estimate the generation probability given an SMC model, which is crucial for the success of our categorization methods. Based on the Markov property defined in (1) and (2), the probability of a sequence of states Q given an SMC model λ, i.e. $P(Q|\lambda)$, can be formatted compactly as

$$P(Q|\lambda) = \pi_{q_{11}} \prod_{1 \le i,j \le M} h_{ij}^{n_{ij}^h(Q)} v_{ij}^{n_{ij}^v(Q)}, \tag{3}$$

where $n_{ij}^h(Q)$ (or $n_{ij}^v(Q)$) is the number of horizontal (or vertical) transitions from state s_i to state s_j in Q.

The second problem in our SMC model is to determine the model parameters to maximize the probability of the sequence given the model. In our case, the estimation problem

can be reduced into a simple maximum likelihood estimation (MLE) of the model parameters. Given a set of 2D sequences $\mathcal{C}$ from one category, the model parameters can then be derived as:

$$\hat{\pi}_i = n_i(\mathcal{C}) / \sum_{i'=1}^{M} n_{i'}(\mathcal{C}), \quad (4)$$

$$\hat{h}_{ij} = n^h_{ij}(\mathcal{C}) / \sum_{j'=1}^{M} n^h_{ij'}(\mathcal{C}), \quad (5)$$

$$\hat{v}_{ij} = n^v_{ij}(\mathcal{C}) / \sum_{j'=1}^{M} n^v_{ij'}(\mathcal{C}), \quad (6)$$

where $n_i(\mathcal{C})$ is the times of state s_i occurring at block $(1,1)$ according to the training set $\mathcal{C}$, $n^h_{ij}(\mathcal{C}) = \sum_{Q\in\mathcal{C}} n^h_{ij}(Q)$, and $n^v_{ij}(\mathcal{C}) = \sum_{Q\in\mathcal{C}} n^v_{ij}(Q)$.

4. Semi-Supervised Categorization

4.1. Spatial Markov Kernel

To develop a graph-based semi-supervised learning approach to image categorization with less labeled data, we then incorporate the above intra-image semantic context into a kind of spatial Markov kernel (SMK) which can be used as the affinity matrix of a graph.

The basic idea of defining a kernel is to map the 2D sequence Q of an image into a high-dimensional feature space: $Q \mapsto \Phi(Q)$. If an SMC model $\lambda^{(Q)}$ is estimated via MLE for each image (i.e. we assume that each "category" has only one sequence Q), the feature mapping Φ can then be given as

$$\Phi(Q) = \lambda^{(Q)} = \{\pi^{(Q)}, H^{(Q)}, V^{(Q)}\}. \quad (7)$$

That is, the sequence Q is now represented by the model parameters of SMC.

Since we focus on capturing the spatial dependencies between states (visual keywords), we only consider the horizontal and vertical transition matrices in $\Phi(Q)$. Moreover, to make the computation more stable, the feature mapping Φ is then defined as

$$\Phi(Q) = \{(h^{(Q)}_{ij} + v^{(Q)}_{ij})/2\}_{1\le i,j\le M}. \quad (8)$$

Though we can map Q into an even higher dimensional feature space by stacking $H^{(Q)}$ and $V^{(Q)}$ in one vector, the experiments show our definition of Φ in (8) is better.

After the feature mapping Φ has been defined, our SMK function in the feature space (determined by Φ) can be given as the following inner-product

$$K(Q, Q') = < \Phi(Q), \Phi(Q') >, \quad (9)$$

where Q and Q' are two sequences. Though the feature vector given by (8) has dimensionality M^2 for M states, the computation of the above kernel is very efficient because the model parameters $H^{(Q)}$ and $V^{(Q)}$ are extremely sparse.

Another advantage of using the kernel method is that different kernels can be readily combined. In our work, the kernel combination is used to deal with the rotation issue that occurs among images from one category with different orientations. That is, we can consider dual SMC models, for which the notions of "past" are defined completely inversely. The obtained two kernels are then combined to make the technique less sensitive to rotation between images of the class. It should be noted that the rotation issue is challenging for HMM and Markov chain model since they are actually directed graphs.

The inverse SMC model can be defined similarly as the original SMC model, and we have to first rewrite the original sequence Q as follows $\tilde{Q} = q_{X,Y} q_{X,Y-1} \cdots q_{X,1} q_{X-1,Y} q_{X-1,Y-1} \cdots q_{X-1,1} \cdots q_{11}$ by an inverse row-wise raster scan on the regular grid. Based on dual SMC models, we then define the feature mapping Φ for SMK as follows:

$$\Phi(Q) = \{(h^{(Q)}_{ij} + v^{(Q)}_{ij} + h^{(\tilde{Q})}_{ij} + v^{(\tilde{Q})}_{ij})/4\}_{1\le i,j\le M}, \quad (10)$$

which ensures rotational invariance with respect to those horizontal or vertical rotations.

Strictly speaking, our SMK method is not rotational invariant, even if dual Markov models are considered. To obtain a rotational invariant solution, we could adopt space-filling curve as the scanning path through all the blocks, without assuming any regional structure of the image. However, it may incur too large a computational cost to make the technique applicable in practice. Fortunately, in natural or histological images the image blocks are usually arranged in an order for normal viewing, and the horizontal or vertical rotations that most possibly occur among the images due to the image preparation process can be handled by our dual Markov models. Hence, the categorization results are shown to be satisfactory in our experiments.

4.2. Label Propagation Using a k-NN Graph

After each image is represented as a sequence of visual keywords, the semi-supervised image categorization problem can be formulated as follows. Given an image data set $\mathcal{Q} = \{Q_1, ..., Q_l, Q_{l+1}, ..., Q_N\}$ and a label set $\mathcal{L} = \{1, ..., C\}$, the first l images $Q_i (i \le l)$ are labeled as $z_i \in \mathcal{L}$ and the remaining images $Q_u (l+1 \le u \le N)$ are unlabeled. The goal is to predict the label of the unlabeled images through label propagation.

Let $\mathcal{F}$ denote the set of $N \times C$ matrices with nonnegative entries. A matrix $F = [F_1^T, ..., F_N^T]^T \in \mathcal{F}$ corresponds to a classification on the image data set $\mathcal{Q}$ by labeling each image Q_i as a label $z_i = \arg\max_{j\le C} F_{ij}$. We can understand F as a vectorial function $F : \mathcal{Q} \to R^C$ which assigns a vector F_i to each image Q_i. Define a $N \times C$ matrix $Z \in \mathcal{F}$

Figure 2. Some sample images from the two image databases: (a) Corel; (b) Histological.

with $Z_{ij} = 1$ if Q_i is labeled as $z_i = j$ and $Z_{ij} = 0$ otherwise. Clearly, Z is consistent with the initial labels according the decision rule. Based on our SMK with incorporated semantic context, the algorithm for label propagation using a k-NN graph is as follows:

(1) Form the affinity matrix W of a k-NN graph by $W_{ij} = \frac{K(Q_i,Q_j)}{\sqrt{K(Q_i,Q_i)}\sqrt{K(Q_j,Q_j)}}$ if Q_j $(j \neq i)$ is among the k-nearest neighbors of Q_i and $W_{ij} = 0$ otherwise (which also implies $W_{ii} = 0$). The distance between Q_i and Q_j is defined as $1 - \frac{K(Q_i,Q_j)}{\sqrt{K(Q_i,Q_i)}\sqrt{K(Q_j,Q_j)}}$, which is used to find the k-nearest neighbors of Q_i.

(2) Construct the matrix $S = D^{-1/2}WD^{-1/2}$ in which D is a diagonal matrix with its (i,i)-element equal to the sum of the i-th row of W.

(3) Iterate $F(t+1) = \alpha SF(t) + (1-\alpha)Z$ for label propagation until convergence, where α is a parameter in the range $(0,1)$.

(4) Let F^* denote the limit of the sequence $\{F(t)\}$. Label each image Q_i as a label $z_i = \arg\max_{j \leq C} F^*_{ij}$.

According to [22], the above algorithm converges to $F^* = (1-\alpha)(I - \alpha S)^{-1}Z$. Since the affinity matrix W is not symmetrical, the prediction result of our algorithm only approximates the solution that minimizes the following cost function associated with F [20]:

$$F^T(I-S)F + \mu(F-Z)^T(F-Z), \tag{11}$$

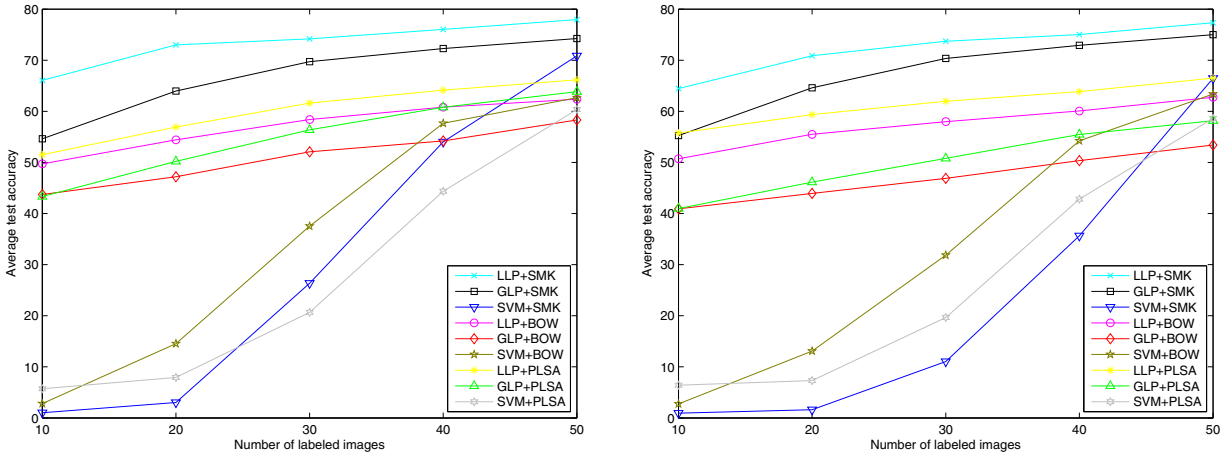
where $\mu > 0$ is the regularization parameter. Our algorithm is different from learning with local and global consistency [22] based on a complete graph. Moreover, though the idea behind our algorithm is also label propagation in neighborhoods [20], we construct the k-NN graph using our spatial Markov kernel, instead of semi-definite quadratic programming which follows the linear neighborhood assumption.

5. Experimental Results

The proposed method for image categorization will be evaluated on the Corel and histological image databases in this section. We first describe the experimental setup, including information of the two image databases and the implementation details. Furthermore, our method is compared with other related works on the two image databases.

5.1. Experimental Setup

The first image database contains 1000 images taken from 10 CD-ROMs published by Corel Corporation. Each CD-ROM contains 100 images representing a distinct category. All the images are of size 384×256 or 256×384.



(a) (b)

Figure 3. The results on the Corel image database as the number of labeled images varies from 10 to 50: (a) $M = 100$; (b) $M = 200$. The average test accuracies are computed on unlabeled data over 10 random runs.

The category names and some randomly selected sample images from each category are shown in Figure 2(a). This is a challenging image database. Some images from two natural scenes (e.g. beach vs. mountains) are difficult to distinguish even by humans.

The second image database is the same as that used in [18, 21], which has five categories (see some samples in Figure 2(b)) of histological images captured from the mentioned five major regions along the human gastrointestinal tract with 40 images for each region. The collection of the histological images is rather time consuming. They were obtained from patient's records in the past 10 years from a local hospital. The image resolution was 4491×3480 pixels during the capturing process. Similar to [18, 21], all the images are then down sampled to 1123×870 pixels.

To form the visual keywords, we have to first divide each image into blocks with the equivalent size $B \times B$ and then extract a joint color/texture feature vector from each block (similar to [21]). The 24 features of this vector are the Gabor textures represented as the means and standard deviations of the coefficients (or outputs) of a bank of Gabor filters [12], which are configured with 3 scales and 4 orientations. Moreover, we also use the color information, i.e., the mean values of the three color components in the HSV color space for the natural images or only the mean gray value for the histological images. Finally, we obtain a 27 or 25 dimensional feature vector for each block, and all the feature vectors are then clustered to form M visual keywords. In the following, we adjust the two parameters B and M with the consideration of the balance of the depiction detail and the computation complexity.

Both supervised and semi-supervised categorization methods are evaluated on the above two databases. For supervised image categorization, the support vector machine (SVM) is used. For semi-supervised image categorization, both complete and k-NN graphs are constructed. In our experiments, we set $\alpha = 0.1$ for the complete graph and $\alpha = 0.9$ for the k-NN graph with $k = 10$. All the other parameters in these methods are set their respective optimal values according to the cross-validation.

5.2. Results of Scene/Object Categorization

We now compare nine methods for image categorization on all the 10 categories of scene/object images from the Corel image database. The notations of three learning algorithms are given as follows:

(1) LLP: semi-supervised categorization using local label propagation on the k-NN graph.

(2) GLP: semi-supervised categorization using global label propagation on the complete graph.

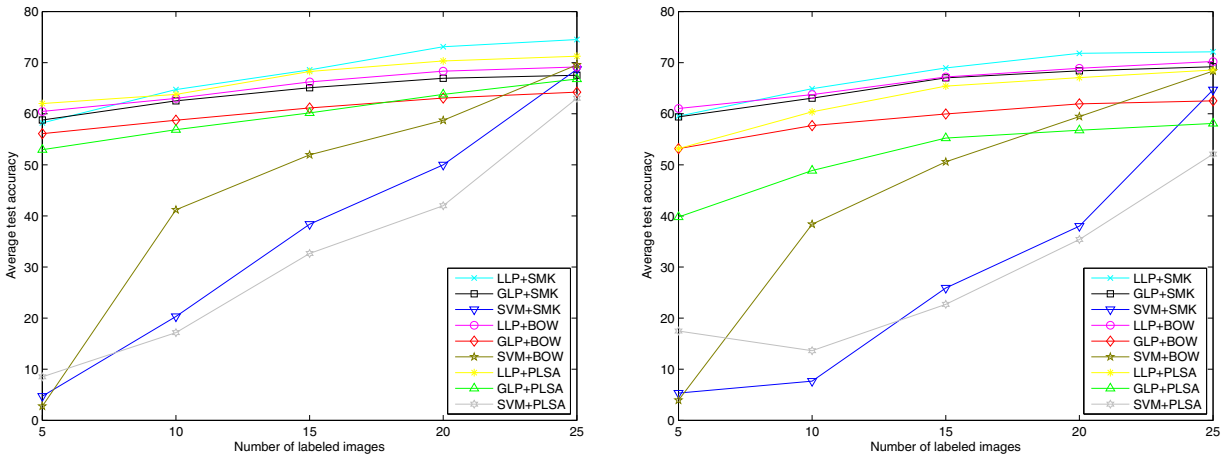
(3) SVM: supervised categorization using the multi-class SVM classifier.

Moreover, we can use three kernels: the proposed spatial Markov kernel (SMK) in this paper, the Gaussian kernel based on bag-of-words (BOW), and the Gaussian kernel based on the learnt topics by PLSA (PLSA). When these kernels are combined with the three learning algorithms, we then have nine methods for image categorization.

In our experiments, we set $B = 16$ and $M = 100$ or 200. The same number of labeled images are randomly selected from each category and the results are averaged over 10 runs. The average test accuracies computed on the unlabeled images over ten categories are then shown in Figure 3 as the number of labeled images varies from 10 to 50. It can be observed that the proposed method (i.e. LLP+SMK)

Methods	skiing	beach	buildings	tigers	owls	elephants	flowers	horses	mountains	food
LLP+SMK	**84.7**	42.3	52.1	**92.4**	88.0	**70.2**	**94.3**	**98.4**	**71.5**	85.5
GLP+SMK	80.6	48.2	59.5	88.2	**89.7**	58.4	84.0	86.8	62.0	85.0
SVM+SMK	72.3	**51.1**	47.8	80.3	80.6	48.2	83.4	88.6	66.8	**89.4**
LLP+BOW	56.2	33.4	39.9	73.5	74.3	60.2	79.0	89.1	41.5	76.8
GLP+BOW	45.4	19.8	60.8	62.2	68.3	55.0	68.4	72.8	45.9	84.5
SVM+BOW	55.4	36.3	46.4	69.7	70.3	49.5	84.3	85.0	58.5	71.4
LLP+PLSA	68.2	36.3	48.6	74.2	74.2	65.6	82.2	89.9	44.5	77.8
GLP+PLSA	59.4	29.2	**63.0**	69.9	70.3	59.3	72.0	82.7	53.4	83.9
SVM+PLSA	60.5	31.6	48.2	66.6	72.5	46.3	72.2	82.2	56.4	66.1

Table 1. The average test accuracies (%) for each category on the Corel image database with 50 labeled images and $M = 100$.



(a) (b)

Figure 4. The results on the histological image database as the number of labeled images varies from 5 to 25: (a) $M = 100$; (b) $M = 200$. The average test accuracies are computed on unlabeled data over 10 random runs.

always outperforms all the other methods. That is, our combination of semantic context and cluster consistency actually leads to significantly better results in scene/object categorization, especially when the number of labeled data is small. Moreover, we can find that the proposed method is not particularly sensitive to the change of the value of M and also that the performance of the semi-supervised methods degrades much more slowly than that of the supervised methods as the number of labeled images becomes smaller. Interestingly, label propagation on a complete graph doesn't even perform better than that on a k-NN graph, due to the fact that the complete graph among all images may map images far away to be nearby in one manifold, causing distortions and hence errors in categorization. As is consistent with the results reported in [14], PLSA doesn't perform better than BOW in all cases. Here, PLSA discovers 80 topics for $M = 100$ and 160 topics for $M = 200$.

In terms of the average test accuracies for each category, it can be found from Table 1 that the proposed method is particularly suitable to process the image content that has an inherent layered structure. For those natural scenes (e.g. skiing, beach, and mountains) that have multiple layers, the proposed method has achieved significant improvements as compared with the methods (i.e. BOW and PLSA) that do not consider the semantic context across visual keywords. Moreover, the proposed method is also shown to be robust despite rotations. That is, when recognizing those animals with different poses (e.g., tigers, elephants, and horses), we can still obtain better results using our SMK.

5.3. Results of Histological Image Categorization

The proposed method is further tested on the histological image database used in [18, 21]. We only consider a fine image representation by dividing each image into small blocks (with respect to the image size 1123×870) with the block size $B = 32$. The nine methods are compared in two cases, i.e. the number of visual keywords is set as $M = 100$ or 200. As for PLSA, the number of latent topics is selected the same as the above section. Moreover, the same number of labeled images are randomly selected from each category and the results are averaged over 10 runs.

The average test accuracies computed on the unlabeled images over five categories are then shown in Figure 4 as the number of labeled images varies from 5 to 25. It can be

observed that the proposed method (i.e. LLP+SMK) generally outperforms all the other eight methods. That is, our combination of semantic context and cluster consistency actually leads to better results in histological image categorization. Moreover, we can find that the proposed method is not particularly sensitive to the change of the number of visual keywords M and also that the performance of the semi-supervised categorization methods degrades significantly more slowly than that of the supervised categorization methods as the number of labeled images becomes smaller. Finally, it can be found that label propagation on a k-NN graph perform better than that on a complete graph for the histological images, which is the same as what has been observed for the Corel images.

6. Conclusions

We have proposed a novel 2D Markov model to capture the spatial dependencies across visual keywords within an image and then avoid the problems with the bag-of-words methods for image categorization. We further incorporate this intra-image semantic context into SMK, for use with graph-based semi-supervised learning to resolve the challenging problem of image categorization with less labeled data. Instead of constructing a complete graph, we resort to a k-NN graph for label propagation with cluster consistency. The categorization experiments on two image databases demonstrate that the proposed method can lead to superior results when we have much less labeled data. In the future work, our SMK will be extended to other applications such as image retrieval and annotation, since the kernel can be regarded as a similarity measure.

Acknowledgements

The work described in this paper was supported by a grant from the Research Council of Hong Kong SAR, China (Project No. CityU 114007) and a grant from City University of Hong Kong (Project No. 7002367).

References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 1, 2

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. COLT*, 1998. 1

[3] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via pLSA. In *Proc. ECCV*, pages 517–530, 2006. 1, 2

[4] P. A. Devijver. Segmentation of binary images using third order Markov mesh image models. In *Proc. ICPR*, pages 259–261, 1986. 1, 3

[5] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, pages 524–531, 2005. 1, 2

[6] P. Franti, O. Virmajoki, and V. Hautamaki. Fast agglomerative clustering using a k-nearest neighbor graph. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1875–1881, 2006. 2

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 41:177–196, 2001. 1, 2

[8] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2):95–114, 2000. 2

[9] S.-S. Kuo and O. E. Agazzi. Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(8):842–848, 1994. 1

[10] J. Li, A. Najmi, and R. Gray. Image classification by a two-dimensional hidden Markov model. *IEEE Trans. on Signal Processing*, 48(2):517–533, 2000. 2

[11] J. Li and J. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, 2003. 2

[12] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996. 6

[13] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proc. CVPR*, 2007. 2

[14] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *Proc. ICCV*, volume 1, pages 883–890, 2005. 1, 2, 7

[15] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 1, 2

[16] F. Salzenstein and C. Collet. Fuzzy Markov random fields versus chains for multispectral image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1753–1767, 2006. 2

[17] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *Proc. ICCV*, pages 42–50, 1998. 2

[18] H. L. Tang, R. Hanka, and H. Ip. Histological image retrieval based on semantic content analysis. *IEEE Trans. on Information Technology in Biomedicine*, 7(1):26–36, 2003. 6, 7

[19] A. Vailaya, A. T. Figueiredo, A. K. Jain, and H. J. Zhang. Image classification for content-based indexing. *IEEE Trans. on Image Processing*, 10(1):117–130, 2001. 2

[20] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *IEEE Trans. on Knowledge and Data Engineering*, 20(1):55–67, 2008. 1, 2, 3, 5

[21] F. Yu and H. Ip. Semantic content analysis and annotation of histological images. *Computers in Biology and Medicine*, 38(6):635–649, 2008. 2, 6, 7

[22] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, 2004. 1, 2, 5