Image Categorization with Spatial Mismatch Kernels

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 new class of 2D string kernels, called spatial mismatch kernels, for use with support vector machine (SVM) in a discriminative approach to the image categorization problem. We first represent images as 2D sequences of those visual keywords obtained by clustering all the blocks that we divide images into on a regular grid. Through decomposing each 2D sequence into two parallel 1D sequences (i.e. the row-wise and column-wise ones), our spatial mismatch kernels can then measure 2D sequence similarity based on shared occurrences of k-length 1D subsequences, counted with up to m mismatches. While those bag-of-words methods ignore the spatial structure of an image, our spatial mismatch kernels can capture the spatial dependencies across visual keywords within the image. Experiments on the natural and histological image databases then demonstrate that our spatial mismatch kernel methods 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 an extremely challenging problem for machines owing to variable and sometimes uncontrolled imaging conditions as well as complex and hard-to-describe objects in an image. In the literature, 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. Another more effective strategy adopts a semantic intermediate representation [3, 4, 15] for each image before image categorization in order to reduce the gap between low-level visual features and high-level semantics and therefore to match the scene/object model with the perception we humans have (e.g. a street scene mainly contains road and buildings), which is then able to classify a larger number of categories.

The bag-of-words methods such as probabilistic latent semantic analysis (PLSA) [8, 9] and latent Dirichlet allocation (LDA) [1, 2] are examples of automatically learning the image semantics using the latter strategy. Each image is first represented by the frequencies of visual keywords, which are learnt through dividing all the images into regions (or blocks) and then applying a clustering algorithm on the visual feature vectors extracted from all the regions (i.e., each cluster is a visual keyword). A mixture of latent topics is further used to model each image, and the latent topics are learnt as a series of multinomial distributions of visual keywords. Although these bag-of-words methods have been shown effective in [3, 4, 15] for natural scene categorization, the semantic context of an image is ignored since all the regions within the image are assumed to be independently drawn from the mixture of latent topics.

However, the natural or histological images usually have a semantic 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. In this paper, to learn the semantic context of the image, we propose a new class of string kernels [12], called spatial mismatch kernels (SMK), for use with support vector machine (SVM) [20] in a discriminative approach to the image categorization problem. The visual keywords are learnt similarly as the bag-of-words methods, but each image has to be divided into blocks on a regular grid so that the spatial position can be characterized for each block. That is, each image can now be represented as a 2D sequence of visual keywords. By decomposing this 2D sequence into two parallel 1D sequences (i.e. the row-wise and column-wise ones), our SMK functions can measure the 2D sequence similarity based on shared occurrences of k-length 1D subsequences, counted with up to m mismatches. This is the first application of mismatch string kernels [11, 23] for 2D sequence matching in image categorization. Here, it should be noted that there are only a small number of keywords (i.e. amino acids) in protein classification that mismatch string kernels were originally applied to, while we usually learn a large number of visual keywords for image categorization and the sequence matching is then much more difficult.

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

Unlike most previous kernel methods that directly modeled the spatial dependencies of low-level visual features extracted from different regions across the image [13, 16] or across different resolutions of the image [5, 6, 10], our SMK methods are applied to image categorization through capturing the spatial dependencies of high level semantic labels (i.e. visual keywords) across an image. Moreover, it should be noted that since no generative model is used in the final categorization our SMK methods are different from those generative methods [7, 21] that also make use of the Markov models to capture the semantic context across visual keywords within an image.

The remainder of this paper is organized as follows. Section 2 gives the definition of the 2D spatial mismatch kernels used to capture the spatial dependencies across visual keywords within an image, and also develops an efficient approach to kernel computation. In Section 3, we give details about the formation of visual keywords used for image representation. In Section 4, our SMK methods are evaluated in the categorization experiments on two image databases. Finally, Section 5 then gives the conclusions drawn from our experimental results.

2. Spatial Mismatch Kernels

To develop a discriminative approach to the image categorization problem, we propose a new class of spatial mismatch kernels (SMK) to capture the semantic context (i.e. spatial dependency) across visual keywords whin an image, for use in categorization with an SVM classifier.

2.1. Definition of SMK

By representing each image as a 2D sequence of visual keywords, we are able to formulate the categorization problem based on the mismatch string kernels. Similar to the bag-of-words methods, all the images are 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. Through clustering all the extracted blocks, a vocabulary of visual keywords $V = \{w_i\}_{i=1}^{M}$ is then generated to exploit the content similarities of blocks. With this universal vocabulary V, we can then represent each image as a 2D sequence Q of visual keywords. In this representation, a visual keyword is automatically attached to each block in the image.

The basic idea of defining a spatial mismatch kernel is to map the 2D sequence Q for an image into a high-dimensional feature space: $Q \mapsto \Phi(Q)$. By decomposing this 2D sequence Q into two parallel 1D sequences, i.e. the row-wise one Q^r and column-wise one Q^c, the feature mapping Φ can be formulated as

$$\Phi(Q) = (\Phi(Q^r)^T, \Phi(Q^c)^T)^T, \tag{1}$$

where $\Phi(Q^r)$ and $\Phi(Q^c)$ are the feature mapping functions for the row-wise and column-wise sequences, respectively. That is, the respective two feature vectors we obtain are stacked together to form a higher dimensional feature vector for the original 2D sequence Q.

More formally, an image Q with $X \times Y$ blocks can now be denoted as a row-wise sequence $Q^r = q_{11}q_{12}...q_{1Y}q_{21}q_{22}...q_{2Y}...q_{XY}$ and a column-wise sequence $Q^c = q_{11}q_{21}...q_{X1}q_{12}q_{22}...q_{X2}...q_{XY}$, where $q_{xy} \in V$ $(1 \le x \le X,\ 1 \le y \le Y)$ is the visual keyword of block (x, y) in the image. In the following, we will only give the details about the feature mapping for the row-wise sequences. The column-wise sequences can be mapped to a high dimensional feature space similarly.

For each k-length subsequence α in a row-wise sequence Q^r (i.e. $\alpha \subset Q^r$), the (k, m)-neighborhood $\mathcal{N}_{(k,m)}(\alpha)$ generated by α is the set of all k-length sequences β from the vocabulary V (i.e. $\beta \in V^k$) that differ from α by at most m mismatches. We then define the following $\Phi_{(k,m)}$ that maps α to a M^k-dimensional feature space:

$$\Phi_{(k,m)}(\alpha) = (\delta_\beta(\alpha))_{\beta \in V^k}, \tag{2}$$

where $\delta_\beta(\alpha) = 1$ if $\beta \in \mathcal{N}_{(k,m)}(\alpha)$, and $\delta_\beta(\alpha) = 0$ otherwise. That is, a k-length subsequence contributes weight to all the coordinates in its mismatch neighborhood.

For a row-wise sequence Q^r, we extend the above feature mapping additively by summing the feature vectors for all the k-length subsequences α in Q^r:

$$\Phi(Q^r) = \sum_{\alpha \in V^k, \alpha \subset Q^r} \Phi_{(k,m)}(\alpha). \tag{3}$$

Note that the β-coordinate of $\Phi(Q^r)$ is a count of all instances of the k-length sequence β occurring with up to m mismatches in Q^r.

Since the feature mapping for the column-wise sequences can be defined similarly, our spatial mismatch kernel can be computed as the following inner-product:

$$\begin{aligned} K(Q_1, Q_2) =& < \Phi(Q_1^r), \Phi(Q_2^r) > \\ &+ < \Phi(Q_1^c), \Phi(Q_2^c) >, \end{aligned} \tag{4}$$

where Q_1 and Q_2 are two 2D sequences (i.e. two images). Since the 2D sequences are mapped to a $2M^k$-dimensional feature space, we do not calculate the feature vectors explicitly but compute their pairwise inner products instead.

2.2. Efficient Computation of SMK

According to the feature mapping defined in (2), the respective feature vectors we obtain are extremely sparse. Therefore, instead of calculating and storing these feature vectors, we directly and efficiently compute the kernel matrix of our SMK for use with an SVM classifier.

Since our SMK function can be decomposed into two components according to (4), without loss of generality, we focus on developing an efficient approach to the kernel computation for the row-wise sequences, which is denoted as $K_r(Q_1, Q_2) =< \Phi(Q_1^r), \Phi(Q_2^r) >$ in the following. Therefore, it follows from (3) that

$$K_r(Q_1, Q_2) = \sum_{\alpha_1 \subset Q_1^r, \alpha_2 \subset Q_2^r} < \Phi_{(k,m)}(\alpha_1), \Phi_{(k,m)}(\alpha_2) >, \quad (5)$$

where both α_1 and α_2 are from V^k.

To obtain the kernel function $K_r(Q_1, Q_2)$, we need compute each $K_{(k,m)}(\alpha_1, \alpha_2) =< \Phi_{(k,m)}(\alpha_1), \Phi_{(k,m)}(\alpha_2) >$ instead, which is a rather difficult task in the M^k-dimensional feature space. Fortunately, this inner-product has the following nice property.

Proposition 1. *For any pair of sequences α_1 and α_2 from V^k, $K_{(k,m)}(\alpha_1, \alpha_2) > 0$ if $\triangle(\alpha_1, \alpha_2) \leq 2m$ and $K_{(k,m)}(\alpha_1, \alpha_2) = 0$ otherwise, where $\triangle(\alpha_1, \alpha_2) = k - \sum_{i=1}^{k} \delta(\alpha_1(i), \alpha_2(i))$ and $\delta(\cdot, \cdot)$ is the Kronecker function.*

Proof. If $\triangle(\alpha_1, \alpha_2) \leq 2m$, these two k-length sequences have at lest $k - 2m$ positions with the same visual keywords. We can construct a k-length sequence $\beta \in V^k$ by concatenating this shared $(k-2m)$-length subsequence α_{12} (maybe not contiguous) and another two m-length subsequences: one from the first m positions of $\alpha_1 - \alpha_{12}$ and the other from the last m positions of $\alpha_2 - \alpha_{12}$. Here, $\alpha_1 - \alpha_{12}$ and $\alpha_2 - \alpha_{12}$ are the $2m$-length subsequences obtained by removing the shared subsequence α_{12} from α_1 and α_2, respectively. The k-length sequence β constructed in this way then satisfies $\triangle(\beta, \alpha_1) \leq m$ and $\triangle(\beta, \alpha_2) \leq m$, i.e., $\beta \in \mathcal{N}_{(k,m)}(\alpha_1)$ and $\beta \in \mathcal{N}_{(k,m)}(\alpha_2)$. Hence, the two feature vectors $\Phi_{(k,m)}(\alpha_1)$ and $\Phi_{(k,m)}(\alpha_2)$ defined by (2) both have nonzero value on one shared coordinate (i.e. β), and we then have $K_{(k,m)}(\alpha_1, \alpha_2) > 0$.

On the other hand, if $K_{(k,m)}(\alpha_1, \alpha_2) > 0$, there exists a k-length sequence β satisfies $\triangle(\beta, \alpha_1) \leq m$ and $\triangle(\beta, \alpha_2) \leq m$, which means that $\triangle(\beta, \alpha_1) + \triangle(\beta, \alpha_2) \leq 2m$. Note that $\triangle(\beta, \alpha_1) + \triangle(\beta, \alpha_2) = 2k - \sum_{i=1}^{k}[\delta(\beta(i), \alpha_1(i)) + \delta(\beta(i), \alpha_2(i))]$. Since it's always true that $\delta(\beta(i), \alpha_1(i)) + \delta(\beta(i), \alpha_2(i)) \leq 1 + \delta(\alpha_1(i), \alpha_2(i))$[1], we have $\triangle(\beta, \alpha_1) + \triangle(\beta, \alpha_2) \geq 2k - \sum_{i=1}^{k}[1 + \delta(\alpha_1(i), \alpha_2(i))] = \triangle(\alpha_1, \alpha_2)$. From $\triangle(\beta, \alpha_1) + \triangle(\beta, \alpha_2) \leq 2m$, it can be followed that $\triangle(\alpha_1, \alpha_2) \leq 2m$. In fact, we have already proven that $\triangle(\alpha_1, \alpha_2) \leq 2m$ is equivalent to $K_{(k,m)}(\alpha_1, \alpha_2) > 0$, which also means that $K_{(k,m)}(\alpha_1, \alpha_2) = 0$ if $\triangle(\alpha_1, \alpha_2) > 2m$. □

[1] If $\beta(i) = \alpha_1(i) = \alpha_2(i)$, $\delta(\beta(i), \alpha_1(i)) + \delta(\beta(i), \alpha_2(i)) = 1 + \delta(\alpha_1(i), \alpha_2(i))$. Otherwise, $\delta(\beta(i), \alpha_1(i)) + \delta(\beta(i), \alpha_2(i)) \leq 1$, and we still have $\delta(\beta(i), \alpha_1(i)) + \delta(\beta(i), \alpha_2(i)) \leq 1 + \delta(\alpha_1(i), \alpha_2(i))$ since $1 + \delta(\alpha_1(i), \alpha_2(i)) \geq 1$.

According to Proposition 1, we only need to compute the inner-product $K_{(k,m)}(\alpha_1, \alpha_2)$ when $\triangle(\alpha_1, \alpha_2) \leq 2m$. As reported in [11], small values of m are most useful to resolve the classification problem. Hence, we focus on considering three cases (i.e., $m = 0, 1, 2$) in this paper and then present the details of $K_{(k,m)}(\alpha_1, \alpha_2)$ as follows:

(1) $m = 0$:

$$K_{(k,m)}(\alpha_1, \alpha_2) = \begin{cases} 1, & \text{if } \triangle(\alpha_1, \alpha_2) = 0; \\ 0, & \text{otherwise.} \end{cases}$$

(2) $m = 1$:

$$K_{(k,m)}(\alpha_1, \alpha_2) = \begin{cases} k(M-1)+1, & \text{if } \triangle(\alpha_1, \alpha_2) = 0; \\ M, & \text{if } \triangle(\alpha_1, \alpha_2) = 1; \\ 2, & \text{if } \triangle(\alpha_1, \alpha_2) = 2; \\ 0, & \text{otherwise.} \end{cases}$$

(3) $m = 2$:

$$K_{(k,m)}(\alpha_1, \alpha_2) = \begin{cases} k(k-1)(M-1)^2/2+1, & \text{if } \triangle(\alpha_1, \alpha_2) = 0; \\ M((k-1)(M-1)+1), & \text{if } \triangle(\alpha_1, \alpha_2) = 1; \\ M^2, & \text{if } \triangle(\alpha_1, \alpha_2) = 2; \\ 6(M-1), & \text{if } \triangle(\alpha_1, \alpha_2) = 3; \\ 6, & \text{if } \triangle(\alpha_1, \alpha_2) = 4; \\ 0, & \text{otherwise.} \end{cases}$$

The above inner-product $K_{(k,m)}(\alpha_1, \alpha_2)$ is computed based on the feature vectors defined in (2) using permutation and combination operations.

We now present the analysis of time complexity for our computation of SMK. For a pair of 2D sequences Q_1 and Q_2 (i.e. two images), the time complexity of computing $K_r(Q_1, Q_2)$ is $O(n^2k)$, where $n = X \times Y$ is the length of a 2D sequence. When there are N images in the database, the computation of the entire kernel matrix has the time complexity $O(N^2n^2k)$. Since the 1D mismatch kernel is computed using a mismatch tree data structure in [11] with the time complexity $O(N^2nk^mM^m)$, our kernel computation is more efficient especially when $M > n$ or $m \geq 2$. Here, it should be noted that we usually set M a large value in the application of image categorization. Moreover, we do not have to construct a complex tree structure, which is especially challenging for 2D sequences.

3. Image Representation with Visual Keywords

Before presenting the experimental results, we will first give the details about the formation of visual keywords used to represent the images in this work. All the images are divided into equivalent blocks on a regular grid, and the spatial position can then be characterized for each block in an

Figure 1. Some sample images from the two image databases: (a) Corel; (b) Histological.

image. That is, each image can now be regarded as a 2D sequence of blocks, which can form the input to our SMK methods. It should be noted that this dense image description is also used in many applications [3, 4], which has been shown to perform better than other methods (e.g. interest points). Intuitively, a dense image description is necessary to capture uniform regions such as sky, calm water, or road surface in many natural scenes.

We further extract a joint color/texture feature vector for each block within an image, similar to [24]. The first 24 dimensions of this feature vector are the Gabor textures represented as the means and standard deviations of the coefficients (or outputs) of a bank of Gabor filters [14], which are configured with 3 scales and 4 orientations. Moreover, we also make use of 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. Hence, we finally obtain a 27 or 25 dimensional feature vector for each block.

Based on this block representation, it is necessary to perform regularization on the block features such that they can be indexed efficiently. Considering that many blocks from different images are very similar in terms of the visual features, we have to group similar blocks through clustering to obtain some representative keywords. In our approach, we create a vocabulary of visual keywords to represent the content of blocks by the k-means method. Each cluster forms a visual keyword. With this universal vocabulary, we can then represent each image as a 2D sequence of visual keywords by a row-wise raster scan on the regular grid.

4. Experimental Results

Our SMK methods for image categorization are 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. Moreover, our SMK methods are compared with other state-of-the-art techniques.

4.1. 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. The ten category names and some randomly selected sample images from each category are shown in Figure 1(a). All the images are of the size 384×256 or 256×384. It should be noted that this is a challenging image database. Some images from two natural scenes (e.g. beach vs. mountains) are difficult to distinguish even by humans.

	$M = 100$				$M = 200$			
	(2,0)-mismatch	(3,1)-mismatch	BOW	PLSA	(2,0)-mismatch	(3,1)-mismatch	BOW	PLSA
skiing	**93.8**	91.8	83.4	84.6	92.4	**92.8**	83.4	82.8
beach	70.2	**70.8**	65.4	63.8	**73.4**	72.6	62.4	64.4
buildings	**87.4**	83.0	76.6	75.2	86.2	**86.4**	77.0	74.6
tigers	**96.0**	**96.0**	93.4	90.8	**96.4**	95.6	94.0	93.6
owls	91.4	**93.6**	89.4	90.0	**91.0**	90.4	88.8	88.0
elephants	**89.2**	83.6	78.8	78.4	82.8	**85.4**	80.6	82.2
flowers	**97.2**	95.0	93.2	92.6	**97.2**	95.6	93.4	93.2
horses	97.0	**97.2**	93.0	91.8	**96.4**	96.0	90.6	90.4
mountains	**83.0**	80.4	69.2	68.8	83.2	**83.8**	70.4	73.2
food	**94.6**	93.0	89.4	89.8	**94.4**	90.6	89.0	89.2
overall	**90.0**	88.4	83.2	82.6	**89.3**	88.9	83.0	83.2
std dev.	1.4	1.6	1.2	1.3	1.2	1.2	0.6	1.6

Table 1. The categorization results (%) on the Corel image database.

The second image database is the same as that used in [18, 24], which has five categories (see some samples from each category in Figure 1(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 those 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 originally set to 4491×3480 pixels during the capturing process. Similar to [18, 24], all the images are then down sampled to 1123×870 pixels.

Since the image classification problem on the above two databases is multi-class, we follow the one-against-one strategy to train an SVM classifier for all the methods that take advantage of SVM. The optimization problem in SVM is resolved by LIBSVM[2], which can adopt the kernel matrix as the input instead of the feature vectors of samples. In our experiments, our SMK is normalized by $K^{norm}(Q_1, Q_2) = \frac{K(Q_1,Q_2)}{\sqrt{K(Q_1,Q_1)}\sqrt{K(Q_2,Q_2)}}$. Different pairs of (k, m) are selected for our SMK, and we find that the results are better when $(k, m) = (2, 0)$ or $(3, 1)$. Moreover, we adopt the Gaussian kernel when the feature vectors of samples are used as the input to the SVM classifier.

To learn visual keywords from each database, there are two parameters (i.e. the block size B and the number of visual keywords M) that should be adjusted considering the balance of the depiction detail and the computation complexity. Actually, we set B a small value to divide each image into blocks of the size $B \times B$ so that we can obtain a fine image representation. Moreover, we set different values to M for clustering on all the block feature vectors with M initial clusters. We aim to test the sensitivity of the performance of our SMK methods to this parameter to provide some empirical results on its choice.

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Moreover, for each database, images within each category are randomly selected into two subsets of the same size to form a training set and a test set, respectively. We repeat each experiment for 10 random splits, and report the average of the results obtained over 10 different test sets. The parameters in SVM are selected according to a twofold cross-validation on the training set. Here, it should be noted that the "optimal" parameters for SVM may change with the number of visual keywords M.

4.2. Results of Scene/Object Categorization

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

(1) (2,0)-mismatch: the SVM classifier using our SMK with $(k, m) = (2, 0)$.

(2) (3,1)-mismatch: the SVM classifier using our SMK with $(k, m) = (3, 1)$.

(3) BOW: the SVM classifier using bag-of-words (i.e. the frequencies of visual keywords used as features).

(4) PLSA: the SVM classifier using the latent topics (i.e. features) learnt by PLSA.

In our experiments, we fix $B = 16$ but try $M = 100$ or 200. The images within each category are randomly selected into two subsets of the same size to form a training set and a test set, respectively. The average categorization accuracies for each category over 10 randomly generated test sets are listed in Table 1, which also gives out the overall accuracies on all the ten categories and the corresponding standard deviations. In terms of the overall accuracies, we can find that our two SMK methods always outperform the

	skiing	beach	buildings	tigers	owls	elephants	flowers	horses	mountains	food
skiing	**93.8**	0.4	3.0	0.2	0.0	0.0	0.0	0.0	2.0	0.6
beach	0.0	**70.2**	6.8	0.0	0.0	2.0	0.0	0.0	17.6	3.4
buildings	0.0	3.8	**87.4**	0.0	0.6	3.4	0.8	0.0	3.0	1.0
tigers	3.0	0.6	0.0	**96.0**	0.4	0.0	0.0	0.0	0.0	0.0
owls	0.0	1.6	2.2	0.0	**91.4**	0.0	2.6	0.0	0.0	2.2
elephants	0.0	1.2	1.6	0.0	0.0	**89.2**	0.0	3.2	3.6	1.2
flowers	0.0	0.0	0.0	0.0	0.0	0.4	**97.2**	1.6	0.2	0.6
horses	0.2	0.2	0.2	0.0	0.0	1.0	0.0	**97.0**	0.2	1.2
mountains	0.2	10.6	0.4	0.0	0.0	4.8	1.0	0.0	**83.0**	0.0
food	0.0	1.0	0.6	0.0	0.0	0.6	1.2	0.0	2.0	**94.6**

Table 2. The confusion matrix (%) for our (2,0)-mismatch kernel method with $M = 100$ on the Corel image database.

Figure 2. Some sample images misclassified between two scenes: beach and mountains. The first row are the beach images misclassified as mountains, while the second row are mountains misclassified as beach.

other two methods. That is, our semantic context analysis of the images with spatial mismatch kernels actually leads to better categorization results. As our two SMK methods are compared, it can be observed that the (2,0)-mismatch kernel method performs a little better than the (3,1)-mismatch kernel method. Moreover, we can also find that our two SMK methods are not particularly sensitive to the change of M as the results are comparable in the two cases. Interestingly, PLSA doesn't perform better than BOW in all the cases, which is also consistent with the results reported in [15]. In our experiments, we let PLSA discover 80 latent topics for $M = 100$ and 160 latent topics for $M = 200$.

In terms of the categorization accuracies for each category, it can be found from Table 1 that our SMK methods perform generally better (or comparably) than the other two methods on all the ten categories, especially when the number of visual keywords M is set a larger value (see the case $M = 200$ in Table 1). Moreover, we can find that our SMK methods are particularly suitable to process those images with a layer structure. For those natural scenes (e.g. beach, buildings, and mountains) that have multiple layers, our SMK methods achieve significant improvements (i.e. about 10%) as compared with those methods (i.e. BOW and PLSA) that do not consider the spatial dependencies across visual keywords within an image.

The confusion matrix for our (2,0)-mismatch kernel method with $M = 100$ is presented in Table 2 to give more details on the categorization of each category. Each row lists the average percentages (over 10 randomly generated test sets) of images in a specific category classified to each of the 10 categories. The numbers on the diagonal show the categorization accuracy for each category and off-diagonal entries indicate categorization errors. A detailed examination of the confusion matrix given by Table 2 shows that six categories, namely skiing, tigers, owls, flowers, horses, and food, obtain the best categorization results (with accuracy $> 90\%$), while two of the largest errors (the underlined numbers in Table 2) arise between two categories: beach and mountains. As for these two natural scenes, 17.6% of beach images are misclassified as mountains while 10.6%

	$M = 100$				$M = 200$			
	(2,0)-mismatch	(3,1)-mismatch	BOW	PLSA	(2,0)-mismatch	(3,1)-mismatch	BOW	PLSA
overall	80.2	**82.7**	80.2	81.6	80.4	**81.7**	78.4	78.0
std dev.	2.0	3.0	3.3	4.3	3.2	3.3	4.2	6.2

Table 3. The categorization results (%) on the histological image database.

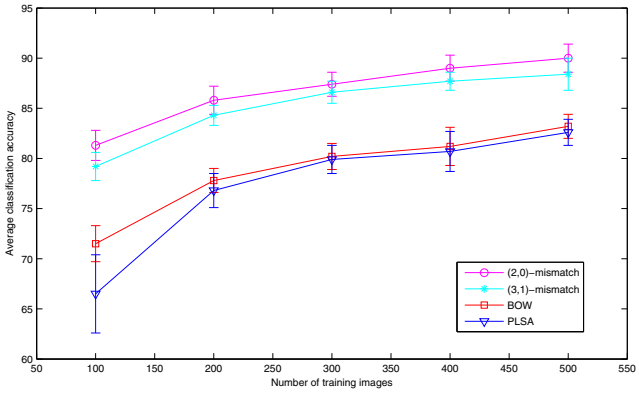

Figure 3. Comparison of the four methods with $M = 100$ on the Corel image database as the number of training images varies from 100 to 500. The average classification accuracies (%) are computed over 10 randomly generated test sets, and the error bars indicate the corresponding standard deviations.

of mountains are misclassified as beach, and some samples are shown in Figure 2. The high errors may be due to the fact that many images from these two scenes have many blocks that are visually similar, such as blocks corresponding to mountains, river, lake, and ocean.

We have also evaluated the four categorization methods with $M = 100$ when there is less training data available. The number of training images has decreased from 500 to 100, and one tenth of the total training images have been selected from each category. The number of test images has correspondingly increased from 500 to 900. As indicated in Figure 3, when the number of training images decreases, the average classification accuracies of our two SMK methods degrade as expected. Moreover, the performances of our two SMK methods are shown to degrade in roughly the same speed as that of BOW. When the number of training images decreases to a small value, we see that the performance of PLSA degrades the fastest. It should be noted that all the supervised learning methods for image categorization will fail when there are much less training data (i.e., each category has only several labeled images), and we have to resort to the semi-supervised methods [22, 25] instead for learning with labeled and unlabeled images.

Finally, the four categorization methods are compared in terms of the computational cost. It should be noted that the computational cost of training SVM is the same for all the four methods if the kernel matrix has been computed first. Hence, we focus on the kernel computation to make a fair comparison. In our experiments, it can be found that BOW obtains the kernel matrix the fastest, our two SMK methods a little slower, and PLSA the slowest due to the costly EM iteration. Here, our SMK methods result in efficient computation of the kernel matrix due to the fact that the feature vectors used to define kernel are extremely sparse.

4.3. Results of Histological Image Categorization

Our SMK methods are further tested on the histological image database which has been used in [18, 24]. 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 four categorization 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 just as the above section. Moreover, the images within each category are randomly selected into two subsets of the same size to form a training set and a test set, respectively. Such random partition is repeated 10 times in our experiments.

The overall accuracies over the five categories and the corresponding standard deviations are used to evaluate the four categorization methods. The average categorization results over the 10 randomly generated test sets are then listed in Table 3. We can find that our (3,1)-mismatch kernel method always outperforms the two methods (i.e. BOW and PLSA) without considering the spatial dependencies across visual keywords within an image. That is, our analysis of the spatial image structure actually leads to better categorization results. Here, it should be noted that our (2,0)-mismatch kernel method outperforms BOW and PLSA only when $M = 200$. Although the (2,0)-mismatch kernel method has been shown to perform better than the (3,1)-mismatch kernel method on the Corel database, this is not true for histological image categorization.

Moreover, it can be observed that our two SMK methods keep robust with respect to the change of M (the same results have been obtained on the Corel image database), while BOW and PLSA can not achieve satisfactory results when M is set a larger value. Therefore, we can find that our (3,1)-mismatch kernel method makes more improvements against BOW and PLSA when $M = 200$.

5. Conclusions

We have introduced a new class of string kernels, i.e. spatial mismatch kernels (SMK), for use with SVM in a discriminative approach to the image categorization problem. We first represent images as 2D sequences of those visual keywords obtained by clustering all the blocks that we divide images into on a regular grid. Through comparing these 2D sequences based on shared occurrences of k-length subsequences counted with up to m mismatches, our SMK function can then capture the spatial structure of an image which is ignored by the bag-of-words methods. The categorization experiments on the Corel and histological image databases then demonstrate that the proposed methods can lead to superior results. In the future work, our SMK methods will be evaluated in other applications such as image retrieval, since our kernels can be regarded as a kind of similarity measures.

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 and M. Jordan. Modeling annotated data. In *Proceedings of the 26th Annual ACM Conference on Research and Development in Information Retrieval*, pages 127–134, 2003. 1

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

[3] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via pLSA. In *Proceedings of European Conference on Computer Vision*, pages 517–530, 2006. 1, 4

[4] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005. 1, 4

[5] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Proceedings of International Conference on Computer Vision*, volume 2, pages 1458–1465, 2005. 2

[6] K. Grauman and T. Darrell. The pyramid match kernel: efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760, 2007. 2

[7] A. Gruber, Y. Weiss, and M. Rosen-Zvi. Hidden topic Markov models. In *Proceedings of the Conference on Artificial Intelligence and Statistics*, 2007. 2

[8] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, 1999. 1

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

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006. 2

[11] C. Leslie, E. Eskin, J. Weston, and W. S. Noble. Mismatch string kernels for SVM protein classification. In *Advances in Neural Information Processing Systems*, volume 15, pages 1417–1424, 2003. 1, 3

[12] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002. 1

[13] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008. 2

[14] 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. 4

[15] 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 *Proceedings of International Conference on Computer Vision*, volume 1, pages 883–890, 2005. 1, 6

[16] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991. 2

[17] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *Proceedings of International Conference on Computer Vision*, pages 42–50, 1998. 1

[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. 5, 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. 1

[20] V. N. Vapnik. An overview of statistical learning theory. *IEEE Trans. on Neural Networks*, 10(5):988–999, 1999. 1

[21] J. Verbeek and B. Triggs. Region classification with Markov field aspect models. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007. 2

[22] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *IEEE Trans. on Knowledge and Data Engineering*, 20(1):55–67, 2008. 7

[23] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. S. Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–3247, 2005. 1

[24] F. Yu and H. Ip. Semantic content analysis and annotation of histological images. *Computers in Biology and Medicine*, 38(6):635–649, 2008. 4, 5, 7

[25] 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. 7